

Dilindungi oleh Undang-Undang Republik Indonesia
Nomor 28 Tahun 2014 tentang Hak Cipta

———— **Rahasia** ————

**MENGIDENTIFIKASI SISWA
BERPRESTASI**

Dengan ————

ALGORITMA NAÏVE BAYES

Sanksi Pelanggaran Pasal 113

Undang-Undang No. 28 Tahun 2014 Tentang Hak Cipta

- i. Setiap Orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf i untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 1 (satu) tahun dan/atau pidana denda paling banyak Rp 100.000.000 (seratus juta rupiah).
- ii. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/atau huruf h untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/atau pidana denda paling banyak Rp 500.000.000,00 (lima ratus juta rupiah).
- iii. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf a, huruf b, huruf e, dan/atau huruf g untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 4 (empat) tahun dan/atau pidana denda paling banyak Rp 1.000.000.000,00 (satu miliar rupiah).
- iv. Setiap Orang yang memenuhi unsur sebagaimana dimaksud pada ayat (3) yang dilakukan dalam bentuk pembajakan, dipidana dengan pidana penjara paling lama 10 (sepuluh) tahun dan/atau pidana denda paling banyak Rp 4.000.000.000,00 (empat miliar rupiah).

Dilindungi oleh Undang-Undang Republik Indonesia
Nomor 28 Tahun 2014 tentang Hak Cipta

Rahasia

MENGIDENTIFIKASI SISWA BERPRESTASI

Dengan

ALGORITMA NAÏVE BAYES

Rulia Pitriani

Angga Putra Juledi, S.Kom., M.Kom

Irmayanti, S.Si., M.pd

Ibnu Rasyid Munthe, S.T., M.Kom



RAHASIA MENGIDENTIFIKASI SISWA BERPRESTASI DENGAN ALGORITMA NAÏVE BAYES

© Rulia Pitriani ; Angga Putra Juledi, S.Kom., M.Kom ;
Irmayanti, S.Si., M.pd; Ibnu Rasyid Munthe, S.T., M.Kom 2026

Pemerhati Aksara — Tim Penerbit Pustaka Riyadz
Perancang Sampul — Tim Penerbit Pustaka Riyadz
Penata Letak— Tim Penerbit Pustaka Riyadz

Cetakan Pertama, April 2026

iv+ 71 hal 15,5 × 23 cm

ISBN : 978-634-7078-80-3

Diterbitkan oleh



Penerbit Pustaka Riyadz
Desa Sawahan Jaya, Semurup, Kec. Air Hangat,
Kerinci - Jambi
pustakariyadz@gmail.com
Instagram @pustakariyadz
HP 0852-6988-7472
Anggota IKAPI : No. 08/JBI/2023

Hak cipta dilindungi undang-undang;

*Dilarang mengutip atau memperbanyak sebagian atau seluruh isi
buku ini tanpa seizin tertulis dari Penerbit*



KATA PENGANTAR

Puji syukur kehadiran Tuhan Yang Maha Esa, karena atas rahmat dan karunia-Nya, buku yang berjudul “Rahasia Mengidentifikasi Siswa Berprestasi dengan Algoritma Naïve Bayes” ini dapat diselesaikan dengan baik.

Buku ini disusun sebagai bentuk kontribusi dalam pengembangan ilmu pengetahuan, khususnya pada bidang data mining dan penerapannya dalam dunia pendidikan. Di era digital saat ini, pengolahan data menjadi salah satu aspek penting dalam pengambilan keputusan yang tepat dan objektif. Salah satu tantangan yang dihadapi institusi pendidikan adalah bagaimana mengidentifikasi siswa berprestasi secara akurat, tidak hanya berdasarkan penilaian subjektif, tetapi juga dengan pendekatan berbasis data.

Melalui buku ini, penulis berusaha menguraikan konsep dasar hingga implementasi algoritma Naïve Bayes sebagai salah satu metode klasifikasi yang sederhana namun efektif. Dengan pendekatan ini, diharapkan proses identifikasi siswa berprestasi dapat dilakukan secara lebih sistematis, terukur, dan berbasis bukti.

Buku ini dirancang agar dapat dipahami oleh berbagai kalangan, mulai dari mahasiswa, peneliti, hingga praktisi pendidikan yang memiliki ketertarikan terhadap pemanfaatan teknologi dalam analisis data. Penulis juga menyajikan contoh kasus dan

pembahasan yang aplikatif untuk memudahkan pembaca dalam memahami penerapan algoritma Naïve Bayes dalam konteks nyata.

ungi oleh Undang-Undang Republik Indonesia

Nomor 28 Tahun 2014 tentang Hak Cipta

Penulis menyadari bahwa buku ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi penyempurnaan di masa mendatang. Akhir kata, penulis berharap buku ini dapat memberikan manfaat, menambah wawasan, serta menjadi referensi yang berguna dalam pengembangan sistem pendukung keputusan di bidang pendidikan.

Penulis

©p.pustakariyadz



KATA PENGANTAR RULIA

Puji syukur kehadiran Allah SWT yang telah melimpahkan rahmat, kesehatan, dan kesempatan kepada penulis sehingga dapat menyelesaikan Buku Monograf ini dengan baik. Shalawat serta salam senantiasa tercurahkan kepada junjungan kita Nabi Besar Muhammad SAW, beserta keluarga, sahabat, dan pengikutnya hingga akhir zaman. Semoga kita semua termasuk ke dalam golongan umat yang mendapatkan syafaatnya di hari akhir kelak. Buku Monograf berjudul: "Rahasia Mengidentifikasi Siswa Berprestasi dengan Algoritma Naïve Bayes".

Dalam proses penyusunan Buku Monograf ini, penulis menyadari bahwa keberhasilan penyusunan tidak terlepas dari dukungan, bantuan, arahan, serta bimbingan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Bapak Alm. Dr. H. Amarullah Nasution, SE., MBA, selaku pendiri Yayasan Universitas Labuhanbatu.
2. Bapak Halomoan, M.H., selaku Ketua Yayasan Universitas Labuhanbatu.
3. Bapak Assoc. Prof. Ade Parlaungan Nasution, PhD, selaku Rektor Universitas Labuhanbatu.
4. Bapak Assoc. Dr. Iwan Purnama, S.Kom., M.Kom, selaku Dekan Fakultas Sains dan Teknologi Universitas Labuhanbatu.



Tidak untuk diperjualbelikan

5. Bapak Budianto Bangun, S.Sos., M.Kom selaku Ketua Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Labuhanbatu
6. Bapak Angga Putra Juledi, S.Kom., M.Kom, selaku Dosen Pembimbing I, yang telah meluangkan waktu, tenaga, dan pikiran dalam memberikan arahan, bimbingan, serta motivasi selama proses penyusunan proposal monograf ini.
7. Ibu Irmayanti ,S.Si., M.pd , selaku Dosen Pembimbing II, yang telah memberikan arahan, masukan, dan bimbingan yang sangat berarti dalam penyempurnaan proposal Monograf ini .
8. Bapak Ibnu Rasyid Munthe,S.T.,M.Kom, selaku Dosen Penguji, yg telah memberi masukan.
9. Kedua orang tua saya Bapak Rusliadi dan Ibu Suwartik beserta adik saya Rendra Agustian Ramadan orang yang hebat yang selalu menjadi penyemangat saya sebagai sandaran terkuat dari keras nya dunia. yang tidak henti-hentinya memberikan kasih sayang dengan penuh cinta dan selalu memberikan motivasi, terimah kasih selalu berjuang untuk kehidupan saya, terimah kasih untuk semuanya berkat do'a dan dukungan bapak Mamak dan Adik saya bisa berada di titik ini. Sehat selalu dan hiduplah lebih lama lagi bapak dan mamak harus selalu ada di setiap perjalanan dan pencapaian hidup saya, Iloveyouyou more more
10. Kepada seseorang yg bernama M.Gio Fanni Pebriansyah terimah kasih telah berjuang bersama dan membersamai penulis selama menyusun tugas akhir dalam kondisis apapun, telah menjadi support sistem dan pendengar keluh kesah penulis.
11. Teman-teman seperjuangan serta semua pihak yang tidak dapat penulis sebutkan satu per satu, yang telah memberikan bantuan, dukungan, dan semangat selama proses penyusunan Buku Monograf ini.
12. Terakhir terimah kasih kepada wanita cantik yaitu diri saya sendiri Rulia Pitriani. Terimah kasih atas perjalanan panjang yang telah di lalui sampai sejauh ini, banyak keluh kesah yang dipendam sendiri, bangga untuk setiap langkah kecilku,

meskipun mudah menangis tapi tidak berhenti mencoba dan menyerah.

Penulis menyadari bahwa Buku Monograf ini masih memiliki keterbatasan dan kekurangan. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun demi kesempurnaan proposal ini.

Labuhanbatu, 2026

Penulis,

RULIA PITRIANI



Tidak untuk diperjualbelikan

Dilindungi oleh Undang-Undang Republik Indonesia
Nomor 28 Tahun 2014 tentang Hak Cipta

©p.pustakariyadz



Tidak untuk diperjualbelikan



DAFTAR ISI

KATA PENGANTAR	i
KATA PENGANTAR RULIA	iii
DAFTAR ISI.....	vii
BAB 01. PENDAHULUAN	1
BAB 02. TEORI DAN KONSEP DATA MINING.....	5
2.1. Data Mining	5
2.2. KDD (Knowledge Discovery in Databases).....	8
2.3. Machine Learning dan Klasifikasi.....	12
2.4. Data Akademik dalam Prediksi Prestasi Siswa.....	15
BAB 03. ALGORITME NAIVE BAYES	19
3.1. Pengertian Naïve Bayes.....	19
3.2. Kelebihan dan Kekurangan.....	21
3.3. Perbandingan dengan Algoritma Lain.....	23
BAB 04. TAHAPAN MACHINE LEARNING	27
4.1. Pengumpulan Data	28
4.2. Pembersihan dan Preprocessing Data.....	29
4.3. Pembagian Data dan Validasi	29
4.4. Pelatihan Model.....	30
4.5. Evaluasi Model.....	30
4.6. Optimasi dan Deployment.....	31
BAB 05. EVALUASI MODEL.....	33
5.1. Metrik Evaluasi	35
5.2. Cross Validation.....	39
5.3. Tantangan Evaluasi	42



Tidak untuk diperjualbelikan

5.4. Perbandingan Kinerja Model.....	44
BAB 06. ALAT BANTU PEMROGRAMAN.....	49
6.1. Penggunaan Python dalam Machine Learning.....	50
6.2. Jupyter Notebook.....	51
BAB 07. PENERAPAN ALGORITMA NAÏVE BAYES	53
7.1. Desain dan Pendekatan Kasus.....	53
7.2. Lokasi dan Subjek Kasus	54
7.3. Prosedur Penerapan Algoritma Naïve Bayes.....	55
7.4. Variabel Penerapan Algoritma Naïve Bayes.....	55
7.5. Teknik Analisis Data Kasus	57
7.6. Pemodelan dan Evaluasi	58
7.7. Penutup.....	59
BAB 08. PENUTUP	61
DAFTAR PUSTAKA	65



Tidak untuk diperjualbelikan



PENDAHULUAN

Pendidikan memiliki peran yang sangat penting dalam membentuk sumber daya manusia yang unggul, berkualitas, dan mampu bersaing di era global yang semakin kompetitif. Melalui pendidikan, individu tidak hanya memperoleh pengetahuan, tetapi juga keterampilan dan pola pikir yang diperlukan untuk menghadapi berbagai tantangan kehidupan. Di tengah perkembangan teknologi dan informasi yang pesat, pendidikan menjadi fondasi utama dalam menciptakan generasi yang adaptif, inovatif, serta mampu berkontribusi secara positif dalam masyarakat.

Pada jenjang sekolah menengah pertama, pendidikan berada pada fase yang sangat krusial dalam perkembangan peserta didik. Pada tahap ini, siswa mulai mengalami transisi dari masa kanak-kanak menuju remaja, sehingga membutuhkan bimbingan yang tidak hanya berorientasi pada aspek akademik, tetapi juga pada pembentukan karakter. Nilai-nilai seperti kejujuran, tanggung jawab, disiplin, dan kerja sama mulai ditanamkan secara lebih intensif agar siswa mampu membangun kepribadian yang kuat dan seimbang.

Selain itu, proses pendidikan di tingkat ini juga berperan dalam membentuk kebiasaan belajar yang baik serta sikap tanggung jawab terhadap tugas dan kewajiban. Lingkungan sekolah menjadi tempat strategis untuk melatih siswa dalam mengelola waktu, menghargai aturan, serta mengembangkan sikap mandiri. Dengan demikian, pendidikan di sekolah menengah pertama tidak hanya bertujuan menghasilkan siswa yang cerdas secara akademik, tetapi juga individu yang berkarakter, disiplin, dan siap menghadapi tantangan di jenjang pendidikan berikutnya maupun dalam kehidupan bermasyarakat.

Seiring dengan meningkatnya kompleksitas sistem pendidikan, sekolah dihadapkan pada tantangan dalam mengelola dan menganalisis data siswa yang semakin besar dari waktu ke waktu. Data akademik seperti nilai, kehadiran, dan partisipasi siswa menjadi komponen penting dalam menilai perkembangan dan potensi siswa. Namun, pengolahan data secara manual sering kali menimbulkan berbagai kendala, seperti keterlambatan dalam pengambilan keputusan, ketidaktepatan analisis, serta adanya unsur subjektivitas.

Perkembangan teknologi informasi memberikan peluang besar untuk mengatasi permasalahan tersebut melalui penerapan data mining dan machine learning. Teknologi ini memungkinkan proses pengolahan data dilakukan secara lebih cepat, akurat, dan objektif dengan cara menemukan pola tersembunyi dari kumpulan data yang tersedia.

Salah satu metode yang cukup populer dan banyak digunakan adalah algoritma Naïve Bayes. Algoritma ini dikenal karena kesederhanaannya serta kemampuannya dalam melakukan klasifikasi berdasarkan pendekatan probabilitas. Dengan memanfaatkan berbagai atribut yang relevan, metode ini mampu membantu dalam mengidentifikasi pola tertentu yang dapat digunakan sebagai dasar pengambilan keputusan.

Penerapan pendekatan berbasis data dalam dunia pendidikan menjadi langkah penting dalam mendukung transformasi digital.

Dengan adanya sistem yang mampu mengolah dan menganalisis data secara efektif, proses evaluasi siswa dapat dilakukan secara lebih objektif dan terukur. Selain itu, hasil analisis juga dapat dimanfaatkan untuk mendukung berbagai kebijakan pendidikan, seperti pengembangan strategi pembelajaran dan peningkatan kualitas pendidikan secara keseluruhan.

Pemanfaatan teknologi dalam dunia pendidikan menjadi langkah strategis untuk menjawab berbagai tantangan pengelolaan data yang semakin kompleks. Dengan semakin banyaknya data akademik yang dihasilkan setiap semester, diperlukan sistem yang mampu mengolah data tersebut secara cepat, akurat, dan terstruktur. Dalam konteks ini, data mining hadir sebagai pendekatan yang memungkinkan penggalian informasi penting dari kumpulan data besar, sehingga dapat membantu pihak sekolah dalam mengambil keputusan yang lebih tepat dan berbasis bukti.

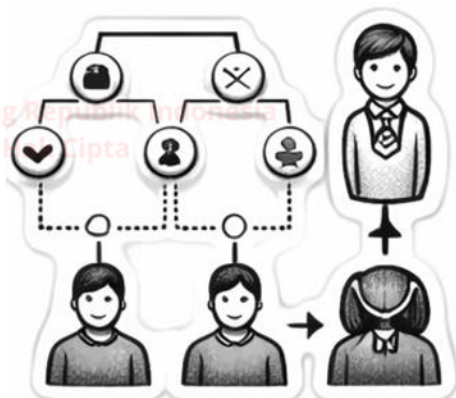
Salah satu metode yang banyak digunakan dalam data mining adalah algoritma Naïve Bayes. Algoritma ini dikenal karena kesederhanaannya dalam perhitungan serta kemampuannya dalam melakukan klasifikasi dengan tingkat akurasi yang cukup baik. Dengan menggunakan prinsip probabilitas, Naïve Bayes dapat mengidentifikasi pola-pola tertentu dari data siswa, seperti hubungan antara nilai akademik, kehadiran, dan aktivitas belajar terhadap prestasi yang dicapai. Hal ini menjadikan algoritma ini sangat relevan untuk diterapkan dalam sistem evaluasi pendidikan.

Melalui penerapan data mining dan algoritma Naïve Bayes, proses identifikasi siswa berprestasi dapat dilakukan secara lebih objektif dan sistematis. Tidak lagi bergantung pada penilaian manual yang rentan terhadap subjektivitas, tetapi berdasarkan analisis data yang terukur. Sistem ini mampu mengelompokkan siswa sesuai dengan tingkat prestasinya, sekaligus memberikan gambaran mengenai faktor-faktor yang memengaruhi keberhasilan belajar. Dengan demikian, guru dan pihak sekolah

dapat lebih mudah dalam merancang strategi pembelajaran yang sesuai dengan kebutuhan masing-masing siswa.

Selain itu, teknologi ini juga memungkinkan deteksi dini terhadap siswa yang berpotensi mengalami penurunan prestasi. Dengan adanya prediksi berbasis data, sekolah dapat melakukan intervensi lebih cepat, seperti memberikan bimbingan tambahan atau pendekatan pembelajaran yang berbeda. Hal ini tidak hanya membantu meningkatkan hasil belajar siswa, tetapi juga mencegah terjadinya kesenjangan prestasi yang semakin lebar di antara peserta didik.

Secara keseluruhan, pemanfaatan data mining dan algoritma Naïve Bayes merupakan solusi inovatif yang dapat meningkatkan efektivitas pengelolaan data pendidikan. Teknologi ini tidak hanya mempercepat proses analisis, tetapi juga menghasilkan informasi yang lebih akurat dan dapat dipertanggungjawabkan. Dengan penerapan yang tepat, diharapkan sekolah dapat mengoptimalkan potensi siswa secara lebih menyeluruh serta meningkatkan kualitas pendidikan secara berkelanjutan di era digital.



TEORI DAN KONSEP DATA MINING

2.1. Data Mining

Data mining merupakan suatu proses yang bertujuan untuk menggali informasi, pengetahuan, atau pola yang tersembunyi dari kumpulan data berukuran besar. Dalam perkembangannya, data mining menjadi bagian penting dalam bidang teknologi informasi karena mampu mengubah data mentah menjadi informasi yang bernilai. Proses ini melibatkan berbagai teknik seperti statistik, machine learning, dan algoritma komputasi untuk menemukan hubungan, tren, maupun pola tertentu yang tidak mudah dikenali secara manual.

Dalam penerapannya, data mining digunakan di berbagai bidang, termasuk pendidikan, bisnis, kesehatan, dan industri. Melalui pendekatan ini, organisasi dapat melakukan analisis prediktif, klasifikasi, clustering, hingga pengambilan keputusan yang lebih akurat dan berbasis data. Misalnya, dalam dunia pendidikan, data mining dapat dimanfaatkan untuk menganalisis prestasi siswa,

mengidentifikasi pola belajar, serta membantu guru dalam menentukan strategi pembelajaran yang lebih efektif.

Dengan demikian, data mining bukan sekadar proses pengolahan data, melainkan suatu pendekatan sistematis untuk mengekstraksi pengetahuan yang bermakna dari data dalam jumlah besar. Kemampuannya dalam mengintegrasikan teknik statistik, machine learning, dan algoritma komputasi menjadikannya sebagai alat yang sangat penting dalam mendukung pengambilan keputusan yang cerdas, cepat, dan berbasis bukti di berbagai bidang kehidupan (Han et al., 2012; Witten et al., 2016; Tan et al., 2019). Dalam konteks machine learning, data mining berfokus pada penemuan pola tersembunyi, hubungan antarvariabel, serta tren dalam data yang dapat digunakan untuk mendukung pengambilan keputusan secara lebih efektif dan akurat (Kotsiantis et al., 2007; Aggarwal, 2015).

Penerapan data mining dalam bidang pendidikan dikenal sebagai Educational Data Mining (EDM), yang bertujuan untuk menganalisis data pendidikan guna memahami perilaku belajar siswa dan meningkatkan kualitas pembelajaran (Romero & Ventura, 2010; Baker & Yacef, 2009). Data mining banyak digunakan untuk memprediksi prestasi akademik siswa melalui pendekatan algoritma klasifikasi, seperti Naïve Bayes, Decision Tree, dan K-Nearest Neighbor (Márquez-Vera et al., 2010; Osmanbegovic & Suljic, 2012). Proses ini melibatkan penggunaan data akademik siswa, seperti nilai mata pelajaran, kehadiran, latar belakang sosial, dan aktivitas pembelajaran, untuk memprediksi hasil akademik siswa, termasuk nilai ujian dan status kelulusan (Cortez & Silva, 2008; Romero et al., 2013).

Melalui teknik data mining, pola dan hubungan yang tidak terlihat secara langsung dalam data dapat diidentifikasi, sehingga memungkinkan institusi pendidikan untuk melakukan intervensi lebih dini terhadap siswa yang berisiko mengalami penurunan prestasi atau gagal lulus (Baker & Inventado, 2014; Peña-Ayala, 2014). Dengan demikian, data mining berperan penting dalam

mendukung pengambilan keputusan berbasis data di bidang pendidikan.

Salah satu metode klasifikasi yang sering digunakan dalam data mining adalah algoritma Naïve Bayes. Algoritma ini digunakan untuk menghitung probabilitas suatu kelas, seperti kategori prestasi atau kelulusan siswa, berdasarkan data historis yang tersedia (Kotsiantis et al., 2007; Han et al., 2012). Algoritma Naïve Bayes merupakan salah satu metode klasifikasi dalam machine learning yang didasarkan pada Teorema Bayes. Ciri utama dari algoritma ini adalah asumsi bahwa setiap atribut atau fitur bersifat independen satu sama lain. Artinya, keberadaan suatu atribut tidak memengaruhi atribut lainnya dalam proses perhitungan probabilitas. Meskipun asumsi ini terkesan sederhana dan tidak selalu sesuai dengan kondisi data di dunia nyata, Naïve Bayes tetap banyak digunakan karena pendekatannya yang praktis dan efisien.

Dalam konteks pendidikan, pemanfaatan algoritma Naïve Bayes menjadi salah satu pendekatan yang efektif dalam memprediksi prestasi akademik siswa. Algoritma ini mampu mengolah berbagai variabel penting seperti nilai akademik, tingkat kehadiran, serta aktivitas belajar untuk menghasilkan klasifikasi yang cukup akurat. Dengan menggabungkan beberapa indikator tersebut, Naïve Bayes dapat memberikan gambaran yang lebih komprehensif mengenai potensi dan capaian siswa dalam proses pembelajaran.

Keunggulan utama dari algoritma ini terletak pada kesederhanaan proses komputasinya. Naïve Bayes tidak memerlukan perhitungan yang rumit maupun sumber daya komputasi yang besar, sehingga sangat mudah diimplementasikan dalam berbagai sistem, termasuk di lingkungan sekolah. Hal ini menjadikannya solusi yang praktis bagi institusi pendidikan yang belum memiliki infrastruktur teknologi yang canggih, namun tetap ingin memanfaatkan analisis data untuk mendukung pengambilan keputusan.

Selain itu, kemampuan Naïve Bayes dalam bekerja dengan dataset berukuran kecil hingga menengah membuatnya sangat relevan untuk diterapkan di sekolah-sekolah dengan keterbatasan data. Algoritma ini tetap mampu memberikan hasil yang optimal meskipun jumlah data tidak terlalu besar. Dengan demikian, Naïve Bayes menjadi pilihan yang tepat untuk membantu sekolah dalam mengidentifikasi prestasi siswa secara lebih objektif, efisien, dan berbasis data, tanpa memerlukan sistem yang kompleks.

Dengan demikian, meskipun Naïve Bayes memiliki keterbatasan berupa asumsi independensi antaratribut, algoritma ini tetap terbukti efektif dalam memprediksi prestasi akademik siswa. Keunggulannya terletak pada kecepatan komputasi, kemudahan implementasi, serta kemampuannya memberikan hasil yang cukup akurat pada berbagai kondisi data. Oleh karena itu, Naïve Bayes menjadi pilihan yang tepat sebagai metode klasifikasi dalam pengembangan sistem prediksi berbasis data di bidang pendidikan (Pandey & Pal, 2011; Wu et al., 2008).

2.2. KDD (Knowledge Discovery in Databases)

KDD (Knowledge Discovery in Databases) merupakan suatu proses sistematis yang bertujuan untuk menemukan pengetahuan yang bernilai dari kumpulan data berukuran besar. Proses ini tidak hanya berfokus pada pengolahan data, tetapi juga mencakup tahapan-tahapan yang terstruktur mulai dari pemilihan data, pembersihan data, transformasi, hingga analisis dan interpretasi hasil. Dengan pendekatan ini, data yang awalnya tidak terstruktur dapat diolah menjadi informasi yang bermakna dan dapat digunakan untuk mendukung pengambilan keputusan.

Dalam praktiknya, KDD melibatkan berbagai teknik seperti statistik, data mining, dan machine learning untuk mengidentifikasi pola, hubungan, serta tren yang tersembunyi dalam data. Proses ini membantu organisasi dalam memahami data secara lebih mendalam, sehingga mampu menghasilkan pengetahuan baru yang sebelumnya tidak terlihat. Misalnya,

dalam dunia pendidikan, KDD dapat digunakan untuk menganalisis performa siswa, mengidentifikasi faktor-faktor yang memengaruhi prestasi, serta membantu merancang strategi pembelajaran yang lebih efektif.

Dengan demikian, KDD tidak hanya sekadar proses analisis data, tetapi merupakan pendekatan komprehensif untuk menghasilkan pengetahuan yang valid, baru, berguna, dan mudah dipahami. Hasil dari proses ini dapat memberikan nilai tambah yang signifikan bagi berbagai bidang, karena mampu mengubah data mentah menjadi informasi yang relevan dan aplikatif. Oleh karena itu, KDD menjadi salah satu komponen penting dalam pengembangan sistem berbasis data di era digital saat ini (Fayyad et al., 1996; Fayyad, 1997). Proses KDD mencakup beberapa tahapan yang saling berkaitan, dimulai dari pemilihan data (*data selection*) dan pembersihan data (*data cleaning*) untuk menghilangkan data yang tidak lengkap, tidak konsisten, atau tidak relevan, sehingga kualitas data yang digunakan dalam analisis dapat terjamin (Han et al., 2012; García et al., 2016). Tahap ini sangat penting karena kualitas data sangat memengaruhi hasil pengetahuan yang dihasilkan (Rahm & Do, 2000).

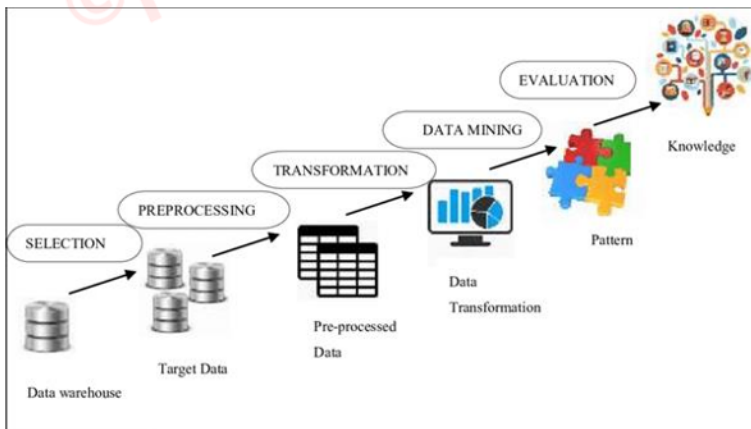
Setelah data dibersihkan, dilakukan tahap transformasi data (*data transformation*), yaitu proses mengubah data ke dalam format atau struktur yang lebih sesuai untuk analisis, seperti normalisasi, agregasi, dan reduksi data (Han et al., 2012; Tan et al., 2019). Transformasi data bertujuan untuk meningkatkan efisiensi serta efektivitas algoritma yang digunakan pada tahap selanjutnya (Witten et al., 2016).

Tahap inti dalam KDD adalah penambangan data (*data mining*), di mana berbagai algoritma statistik dan machine learning diterapkan untuk mengidentifikasi pola, hubungan, atau model tertentu dalam data (Fayyad et al., 1996; Witten et al., 2016; Aggarwal, 2015). Algoritma yang umum digunakan pada tahap ini meliputi klasifikasi, klustering, asosiasi, dan prediksi, tergantung

pada tujuan analisis yang ingin dicapai (Han et al., 2012; Wu et al., 2008).

Hasil dari proses data mining kemudian dievaluasi dan diinterpretasikan untuk menentukan tingkat akurasi, relevansi, serta kegunaannya dalam konteks permasalahan yang diteliti (Fayyad et al., 1996; Tan et al., 2019). Tahap evaluasi ini bertujuan untuk memastikan bahwa pola atau pengetahuan yang dihasilkan benar-benar memiliki makna dan dapat diterapkan dalam pengambilan keputusan (García et al., 2016). Pada tahap akhir, pengetahuan yang telah divalidasi disajikan dalam bentuk yang mudah dipahami, seperti visualisasi, laporan, atau model prediktif, sehingga dapat digunakan untuk mendukung perumusan strategi dan pengambilan keputusan (Han et al., 2012; Witten et al., 2016).

KDD memiliki peran yang sangat penting dalam berbagai bidang, termasuk data mining dan machine learning, khususnya dalam pengolahan data pendidikan. Dalam konteks *Educational Data Mining*, KDD digunakan untuk mengungkap pola tersembunyi dalam data akademik siswa, seperti perilaku belajar, tingkat kelulusan, dan prediksi prestasi akademik siswa, yang tidak dapat diamati secara langsung melalui analisis konvensional (Romero & Ventura, 2010; Baker & Yacef, 2009; Peña-Ayala, 2014).



Gambar 1. Tahapan Proses KDD

Berikut adalah penjelasan mendetail mengenai tahapan-tahapan dalam *Knowledge Discovery In Database* (KDD):

1. Seleksi atau Pemilihan Data

Pada fase ini, data yang relevan untuk analisis diidentifikasi dan dipilih. Pemilihan data yang tepat sangatlah penting karena kualitas dan relevansi data mempengaruhi hasil akhir dari proses KDD.

2. Pemrosesan dan Pembersihan Data

Pemrosesan dan pembersihan data adalah dua tahapan krusial dalam siklus *Knowledge Discovery In Database* (KDD). Proses ini bertujuan untuk memastikan bahwa data yang digunakan dalam analisis atau pemodelan tidak hanya lengkap dan konsisten, tetapi juga sesuai dan siap digunakan dalam model analisis. Tanpa pemrosesan dan pembersihan data yang tepat, hasil analisis atau model yang dibangun mungkin menyesatkan atau tidak valid.

3. Transformasi

Setelah data dibersihkan, Langkah selanjutnya adalah mentransformasikan data. Transformasi data bertujuan untuk mengubah data yang telah diseleksi dan dibersihkan kedalam format atau bentuk yang lebih sesuai untuk dianalisis lebih lanjut. Pada tahap ini, menggunakan Teknik tertentu, seperti data mining.

4. Data Mining

Data mining adalah langkah utama KDD. Pada fase ini diterapkan teknik analisis algoritma untuk menemukan pola atau informasi tersembunyi dalam data. Data mining melibatkan penggunaan berbagai teknik statistik dan pembelajaran mesin untuk menganalisis data dan menemukan hubungan atau pola yang sebelumnya tidak diketahui.

5. Evaluasi

Setelah pola atau model berhasil ditemukan dalam proses analisis data, langkah selanjutnya adalah tahap evaluasi. Tahap ini merupakan bagian penting untuk memastikan bahwa model yang dihasilkan benar-benar mampu

merepresentasikan data dengan baik. Evaluasi dilakukan dengan cara menguji model terhadap data yang ada untuk melihat sejauh mana model tersebut mampu menghasilkan prediksi atau klasifikasi yang sesuai dengan tujuan yang diinginkan.

Pada fase ini, berbagai metode pengujian dapat digunakan untuk mengukur kinerja model, seperti tingkat akurasi, presisi, recall, maupun metrik evaluasi lainnya yang relevan. Hasil evaluasi ini akan menunjukkan apakah model yang dibangun sudah cukup baik atau masih memerlukan perbaikan. Jika performa model belum memenuhi standar yang diharapkan, maka proses sebelumnya seperti pemilihan data atau metode analisis dapat diulang untuk memperoleh hasil yang lebih optimal.

Dengan demikian, evaluasi bertujuan untuk memastikan bahwa pola atau model yang ditemukan memiliki tingkat validitas yang tinggi dan dapat dipercaya. Model yang telah lolos tahap evaluasi dapat digunakan sebagai dasar dalam pengambilan keputusan, karena telah terbukti mampu memberikan hasil yang relevan dan akurat sesuai dengan kebutuhan analisis.

2.3. Machine Learning dan Klasifikasi

Klasifikasi merupakan salah satu teknik utama dalam machine learning yang digunakan untuk mengelompokkan data ke dalam kategori atau kelas tertentu berdasarkan pola yang dipelajari dari data latih. Dalam proses ini, algoritma akan mempelajari hubungan antara atribut-atribut yang ada pada data dengan label kelas yang telah ditentukan sebelumnya. Hasil dari pembelajaran tersebut kemudian digunakan untuk memprediksi kelas dari data baru yang belum pernah dilihat sebelumnya.

Dalam praktiknya, teknik klasifikasi banyak digunakan dalam berbagai bidang, seperti pendidikan, kesehatan, bisnis, dan teknologi informasi. Misalnya, dalam dunia pendidikan, klasifikasi dapat digunakan untuk mengelompokkan siswa berdasarkan tingkat prestasi, seperti berprestasi atau tidak berprestasi. Proses

ini membantu pihak sekolah dalam memahami karakteristik siswa serta menentukan strategi pembelajaran yang lebih tepat. Beberapa algoritma yang sering digunakan dalam klasifikasi antara lain Naïve Bayes, Decision Tree, dan K-Nearest Neighbor.

Dengan demikian, klasifikasi tidak hanya berfungsi sebagai alat pengelompokan data, tetapi juga sebagai dasar dalam pengambilan keputusan berbasis data. Kemampuannya dalam mengenali pola dari data latih menjadikannya metode yang efektif untuk memprediksi dan mengkategorikan data secara otomatis. Oleh karena itu, teknik klasifikasi menjadi komponen penting dalam pengembangan sistem cerdas yang mampu memberikan solusi yang akurat dan efisien di berbagai bidang. (Han et al., 2012; Witten et al., 2016; Aggarwal, 2015). Dalam konteks prediksi prestasi akademik siswa tingkat SMP, metode klasifikasi digunakan untuk memprediksi hasil akademik siswa, seperti kelulusan, kategori nilai, atau tingkat risiko akademik, berdasarkan data historis yang tersedia (Cortez & Silva, 2008; Romero & Ventura, 2010). Data akademik yang umum digunakan dalam proses klasifikasi meliputi nilai ujian sebelumnya, tingkat kehadiran, latar belakang akademik, serta atribut mata pelajaran atau program studi siswa (Yağcı, 2022; Khairy et al., 2024; Osmanbegovic & Suljic, 2012).

Salah satu algoritma klasifikasi yang banyak digunakan dalam prediksi prestasi akademik adalah Naïve Bayes. Naïve Bayes merupakan model probabilistik yang didasarkan pada Teorema Bayes dengan asumsi bahwa setiap fitur bersifat independen secara bersyarat terhadap fitur lainnya (Mitchell, 1997; Han et al., 2012). Meskipun asumsi independensi ini sering kali tidak sepenuhnya terpenuhi dalam data nyata, Naïve Bayes tetap menunjukkan kinerja yang baik dalam berbagai permasalahan klasifikasi, termasuk pada bidang pendidikan (Gandey & Wal, 2011; Wu et al., 2008). Popularitas algoritma ini disebabkan oleh kemudahan implementasi, efisiensi komputasi, serta kemampuannya dalam menangani dataset berukuran kecil hingga menengah (Zheng & Li, 2024; Kotsiantis et al., 2007). Dalam

penelitian prediksi prestasi akademik, Naïve Bayes digunakan untuk menghitung probabilitas kelulusan atau kategori nilai siswa berdasarkan atribut akademik yang tersedia.

Untuk menilai kinerja model klasifikasi, diperlukan proses evaluasi yang sistematis menggunakan berbagai metrik performa. Evaluasi ini bertujuan untuk mengetahui seberapa baik model dalam mengklasifikasikan data sesuai dengan label yang sebenarnya. Tanpa evaluasi yang tepat, model yang dihasilkan belum tentu dapat diandalkan dalam pengambilan keputusan, meskipun terlihat bekerja dengan baik pada data tertentu.

Beberapa metrik yang umum digunakan dalam evaluasi model klasifikasi antara lain akurasi, precision, recall, dan F1-score. Akurasi mengukur tingkat ketepatan keseluruhan prediksi model, yaitu perbandingan antara jumlah prediksi benar dengan total data. Precision menunjukkan seberapa tepat model dalam mengidentifikasi kelas positif, sedangkan recall mengukur kemampuan model dalam menemukan seluruh data yang benar-benar termasuk dalam kelas positif. F1-score merupakan kombinasi dari precision dan recall yang memberikan gambaran keseimbangan antara keduanya, terutama ketika data tidak seimbang.

Selain itu, ROC-AUC juga menjadi metrik penting dalam mengevaluasi performa model klasifikasi. ROC (Receiver Operating Characteristic) menggambarkan hubungan antara tingkat true positive rate dan false positive rate, sedangkan AUC (Area Under Curve) menunjukkan seberapa baik model dalam membedakan antara kelas positif dan negatif. Semakin tinggi nilai AUC, semakin baik kemampuan model dalam melakukan klasifikasi. Dengan menggunakan berbagai metrik ini secara bersamaan, evaluasi model menjadi lebih komprehensif dan dapat memberikan gambaran yang lebih akurat mengenai kinerja model (Powers, 2011; Fawcett, 2006).

Tidak untuk diperjualbelikan

Metrik evaluasi digunakan untuk mengukur sejauh mana suatu model klasifikasi mampu bekerja dengan baik dalam

mengelompokkan data secara tepat. Melalui metrik ini, dapat diketahui tingkat keberhasilan model dalam memprediksi kelas yang benar serta mengidentifikasi kesalahan yang terjadi selama proses klasifikasi. Evaluasi ini menjadi sangat penting karena model yang tampak baik secara umum belum tentu memiliki performa yang optimal pada semua kondisi data.

Dalam konteks pendidikan, penggunaan metrik evaluasi menjadi semakin krusial karena hasil klasifikasi dapat memengaruhi keputusan akademik siswa. Misalnya, kesalahan dalam mengklasifikasikan siswa berprestasi atau tidak berprestasi dapat berdampak pada pemberian penghargaan, penempatan program pembelajaran, atau intervensi pendidikan. Oleh karena itu, model harus mampu meminimalkan kesalahan prediksi agar keputusan yang diambil benar-benar mencerminkan kondisi siswa secara akurat dan adil.

Dengan demikian, penerapan metrik evaluasi tidak hanya berfungsi sebagai alat ukur teknis, tetapi juga sebagai dasar dalam memastikan kualitas dan keandalan model yang digunakan. Evaluasi yang baik akan membantu menghasilkan model yang lebih akurat, objektif, dan dapat dipertanggungjawabkan, sehingga dapat mendukung pengambilan keputusan yang tepat dalam dunia pendidikan serta meningkatkan kualitas proses pembelajaran secara keseluruhan (Yağcı, 2022; Agrawal et al., 2025; Sokolova & Lapalme, 2009).

2.4. Data Akademik dalam Prediksi Prestasi Siswa

Data akademik memiliki peran yang sangat penting dalam memprediksi prestasi akademik siswa karena mencerminkan berbagai aspek yang berkaitan dengan proses belajar. Data seperti nilai ujian, tugas, dan hasil evaluasi lainnya dapat memberikan gambaran mengenai kemampuan kognitif siswa. Selain itu, data akademik juga sering mencakup informasi tentang kehadiran, partisipasi di kelas, serta keterlibatan dalam kegiatan sekolah, yang semuanya berkontribusi terhadap keberhasilan belajar.

Melalui analisis data akademik, pihak sekolah dapat memahami pola belajar siswa secara lebih mendalam. Misalnya, siswa dengan kehadiran tinggi dan partisipasi aktif cenderung memiliki prestasi yang lebih baik dibandingkan siswa yang kurang terlibat. Dengan memanfaatkan data ini, guru dapat mengidentifikasi siswa yang membutuhkan perhatian lebih awal, sehingga dapat diberikan bimbingan atau strategi pembelajaran yang sesuai dengan kebutuhan masing-masing.

Dengan demikian, data akademik tidak hanya berfungsi sebagai alat penilaian, tetapi juga sebagai dasar dalam pengambilan keputusan yang lebih tepat dan objektif. Pemanfaatan data secara optimal memungkinkan sekolah untuk memprediksi prestasi siswa secara lebih akurat, sekaligus meningkatkan kualitas proses pembelajaran. Hal ini sangat penting untuk mendukung perkembangan siswa secara menyeluruh, baik dari segi akademik maupun non-akademik (Cortez & Silva, 2008; Romero & Ventura, 2010).

Variabel merupakan komponen penting dalam penelitian prediksi prestasi akademik siswa karena menjadi dasar dalam proses analisis dan pemodelan. Variabel yang umum digunakan meliputi nilai mata pelajaran sebelumnya, tingkat kehadiran, serta berbagai atribut lain yang berkaitan dengan aktivitas belajar siswa. Nilai akademik mencerminkan kemampuan kognitif siswa, sementara kehadiran menunjukkan tingkat kedisiplinan dan keterlibatan dalam proses pembelajaran di kelas.

Selain itu, variabel seperti partisipasi kelas juga memiliki peran penting dalam menggambarkan keaktifan siswa selama kegiatan belajar berlangsung. Siswa yang aktif bertanya, berdiskusi, dan berinteraksi dengan guru cenderung memiliki pemahaman yang lebih baik terhadap materi. Beban mata pelajaran juga menjadi faktor yang perlu diperhatikan, karena jumlah dan tingkat kesulitan mata pelajaran dapat memengaruhi kemampuan siswa dalam mengelola waktu dan menyerap materi pembelajaran secara optimal.

Dengan menggabungkan berbagai variabel tersebut, model prediksi prestasi akademik dapat menghasilkan analisis yang lebih komprehensif dan akurat. Setiap variabel memberikan kontribusi yang saling melengkapi dalam menggambarkan kondisi belajar siswa secara menyeluruh. Oleh karena itu, pemilihan variabel yang tepat sangat penting untuk memastikan bahwa hasil prediksi dapat digunakan sebagai dasar dalam pengambilan keputusan pendidikan yang lebih efektif dan objektif (Han et al., 2012; Osmanbegovic & Suljic, 2012). Nilai mata pelajaran sebelumnya sering dijadikan sebagai prediktor utama dalam memprediksi prestasi akademik di masa mendatang karena mencerminkan penguasaan materi dan konsistensi belajar siswa (Pandey & Pal, 2011; Khairy et al., 2024).

Selain nilai akademik, kehadiran siswa juga merupakan faktor penting dalam proses prediksi prestasi akademik. Tingkat kehadiran yang tinggi menunjukkan keterlibatan aktif siswa dalam kegiatan pembelajaran, yang secara signifikan berkorelasi dengan hasil akademik yang lebih baik (Romero et al., 2013; Yağcı, 2022). Sebaliknya, tingkat ketidakhadiran yang tinggi sering dikaitkan dengan penurunan prestasi akademik dan meningkatnya risiko ketidakihtutan siswa (Baker & Yacef, 2009; Márquez-Vera et al., 2010).

Berbagai penelitian terdahulu menunjukkan bahwa algoritma Naïve Bayes memiliki kemampuan yang cukup efektif dalam memprediksi prestasi akademik siswa, termasuk nilai mata pelajaran maupun status kelulusan di tingkat sekolah menengah pertama. Hal ini dimungkinkan karena algoritma tersebut mampu mengolah data akademik historis, seperti nilai sebelumnya dan tingkat kehadiran, untuk menemukan pola yang berkaitan dengan keberhasilan belajar siswa. Dengan pendekatan probabilistik, Naïve Bayes dapat memberikan prediksi yang cukup akurat meskipun menggunakan model yang relatif sederhana.

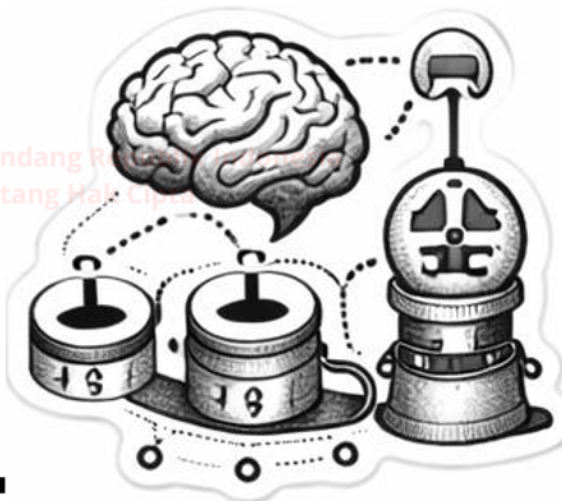
Efektivitas algoritma ini juga didukung oleh kemampuannya dalam menangani data dengan jumlah terbatas, yang sering

menjadi kondisi umum di lingkungan sekolah. Data historis seperti nilai raport dan kehadiran dapat digunakan sebagai indikator utama dalam menentukan kecenderungan prestasi siswa di masa depan. Dengan memanfaatkan data tersebut, model dapat membantu mengidentifikasi siswa yang berpotensi lulus dengan baik maupun siswa yang membutuhkan perhatian lebih dalam proses pembelajaran.

Dengan demikian, penerapan algoritma Naïve Bayes dalam penelitian prediksi akademik memberikan kontribusi yang signifikan dalam meningkatkan kualitas evaluasi pendidikan. Model ini tidak hanya membantu dalam menghasilkan prediksi yang objektif dan berbasis data, tetapi juga mendukung pengambilan keputusan yang lebih tepat oleh pihak sekolah. Oleh karena itu, Naïve Bayes menjadi salah satu metode yang relevan dan praktis untuk diterapkan dalam sistem prediksi prestasi siswa di tingkat SMP (Pandey & Pal, 2011; Yağcı, 2022; Zheng & Li, 2024).

Meskipun Naïve Bayes menggunakan asumsi independensi antar fitur, algoritma ini tetap mampu memberikan hasil prediksi yang cukup akurat serta efisien secara komputasi, sehingga cocok diterapkan pada sistem pendukung keputusan di bidang pendidikan (Han et al., 2012; Wu et al., 2008). Dengan demikian, pemanfaatan data akademik historis melalui algoritma Naïve Bayes dapat membantu pihak sekolah dalam memprediksi prestasi siswa dan melakukan intervensi akademik secara lebih dini dan tepat sasaran.

ALGORITME NAÏVE BAYES



3.1. Pengertian Naïve Bayes

Algoritma Naïve Bayes merupakan salah satu metode klasifikasi yang bekerja berdasarkan Teorema Bayes, yaitu suatu pendekatan probabilistik untuk menghitung kemungkinan suatu kejadian berdasarkan informasi sebelumnya. Dalam konteks machine learning, algoritma ini digunakan untuk menentukan probabilitas suatu data termasuk ke dalam kelas tertentu berdasarkan fitur-fitur yang dimilikinya. Dengan memanfaatkan prinsip ini, Naïve Bayes mampu mengolah data secara efisien untuk menghasilkan prediksi yang cukup akurat.

Dalam proses kerjanya, Naïve Bayes menghitung probabilitas posterior, yaitu probabilitas suatu kelas setelah mempertimbangkan data atau fitur yang tersedia. Model ini akan memilih kelas dengan nilai probabilitas posterior tertinggi sebagai hasil prediksi. Perhitungan ini melibatkan dua komponen utama, yaitu probabilitas prior yang menggambarkan kemungkinan awal

suatu kelas, serta probabilitas likelihood yang menunjukkan hubungan antara fitur dengan kelas tersebut.

Keunggulan utama dari pendekatan ini terletak pada asumsi independensi antar fitur. Artinya, setiap fitur dianggap tidak saling bergantung satu sama lain dalam memengaruhi hasil klasifikasi. Meskipun asumsi ini tidak selalu sesuai dengan kondisi nyata, dalam banyak kasus Naïve Bayes tetap mampu memberikan hasil yang baik. Hal ini menjadikan algoritma ini sederhana, cepat, dan mudah diterapkan dalam berbagai bidang, termasuk pendidikan.

Dalam konteks prediksi prestasi akademik siswa, fitur yang digunakan dalam model Naïve Bayes dapat mencakup berbagai variabel yang relevan. Beberapa di antaranya adalah nilai mata pelajaran sebelumnya, tingkat kehadiran, partisipasi dalam kegiatan belajar, serta atribut lain yang berkaitan dengan kinerja akademik siswa. Dengan menggabungkan berbagai fitur tersebut, model dapat memberikan gambaran yang lebih komprehensif mengenai kemungkinan prestasi siswa di masa depan.

Dengan demikian, Naïve Bayes menjadi metode yang efektif dalam mengklasifikasikan dan memprediksi data akademik siswa. Kemampuannya dalam menghitung probabilitas secara sistematis dan memilih kelas dengan peluang tertinggi menjadikannya alat yang berguna dalam mendukung pengambilan keputusan berbasis data. Penerapan algoritma ini diharapkan dapat membantu pihak sekolah dalam memahami potensi siswa secara lebih objektif dan meningkatkan kualitas proses evaluasi pendidikan (Perez & Perez, 2021; Khairy et al., 2024).

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Algoritma Naïve Bayes didasarkan pada asumsi bahwa setiap fitur atau atribut bersifat independen satu sama lain ketika kelas sudah diketahui. Dalam konteks pendidikan, fitur-fitur tersebut dapat

berupa nilai mata pelajaran, tingkat kehadiran, maupun indikator lain yang berkaitan dengan kinerja siswa. Artinya, setiap variabel dianggap memiliki kontribusi tersendiri dalam menentukan hasil klasifikasi tanpa dipengaruhi oleh variabel lain. Meskipun asumsi ini terkesan sederhana, pendekatan ini justru memudahkan proses perhitungan dan implementasi model.

Pada kenyataannya, hubungan antar fitur sering kali saling berkaitan, misalnya siswa dengan kehadiran tinggi cenderung memiliki nilai yang lebih baik. Namun, meskipun asumsi independensi tidak sepenuhnya mencerminkan kondisi nyata, Naïve Bayes tetap mampu menghasilkan performa yang baik dalam banyak kasus. Hal ini menunjukkan bahwa algoritma ini cukup robust dan dapat digunakan secara efektif, terutama pada data dengan kompleksitas yang tidak terlalu tinggi seperti data akademik siswa.

Selain itu, Naïve Bayes menghasilkan nilai probabilitas untuk setiap kelas yang mungkin, sehingga memungkinkan sistem untuk memilih kategori dengan probabilitas tertinggi sebagai hasil prediksi. Dalam konteks pendidikan, hal ini sangat membantu dalam menentukan prediksi nilai mata pelajaran atau status kelulusan siswa berdasarkan data yang tersedia. Dengan demikian, algoritma ini tidak hanya memberikan hasil klasifikasi, tetapi juga memberikan dasar probabilistik yang dapat digunakan untuk mendukung pengambilan keputusan yang lebih objektif dan terukur.

3.2. Kelebihan dan Kekurangan

3.2.1. Kelebihan

Algoritma Naïve Bayes memiliki berbagai keuntungan yang menjadikannya salah satu metode klasifikasi yang populer dalam machine learning. Salah satu keunggulan utamanya adalah kesederhanaan dalam konsep dan implementasi. Algoritma ini tidak memerlukan proses komputasi yang kompleks, sehingga dapat dengan mudah diterapkan bahkan oleh pengguna yang

memiliki pemahaman dasar tentang analisis data. Selain itu, waktu pelatihan dan pengujian model relatif cepat, sehingga sangat efisien digunakan dalam berbagai aplikasi.

Keuntungan lain dari Naïve Bayes adalah kemampuannya dalam menangani dataset dengan ukuran kecil hingga menengah tanpa mengurangi tingkat akurasi secara signifikan. Algoritma ini juga cukup fleksibel dalam mengolah berbagai jenis data, baik numerik maupun kategorikal. Selain itu, Naïve Bayes mampu bekerja dengan baik meskipun data yang digunakan tidak sepenuhnya sempurna, seperti adanya data yang tidak lengkap atau memiliki noise, sehingga tetap memberikan hasil yang stabil.

Selain efisiensi dan fleksibilitas, Naïve Bayes juga memberikan output dalam bentuk probabilitas untuk setiap kelas, sehingga hasil prediksi dapat lebih mudah dipahami dan dianalisis. Hal ini sangat membantu dalam pengambilan keputusan, karena pengguna tidak hanya mengetahui hasil klasifikasi, tetapi juga tingkat keyakinan dari prediksi tersebut. Dengan berbagai keunggulan tersebut, Naïve Bayes menjadi pilihan yang tepat untuk berbagai kebutuhan analisis data, termasuk dalam bidang pendidikan untuk memprediksi prestasi akademik siswa.

3.2.2. Kekurangan

Algoritma Naïve Bayes memiliki beberapa kekurangan yang perlu dipertimbangkan sebelum digunakan dalam suatu penelitian atau aplikasi. Salah satu kelemahan utamanya adalah asumsi independensi antar fitur. Algoritma ini menganggap bahwa setiap atribut tidak saling berkaitan, padahal dalam kenyataannya banyak variabel yang memiliki hubungan erat. Misalnya, dalam konteks pendidikan, nilai siswa sering dipengaruhi oleh kehadiran dan partisipasi kelas. Asumsi yang terlalu sederhana ini dapat menyebabkan hasil prediksi menjadi kurang akurat pada data yang kompleks.

Kekurangan lainnya adalah ketergantungan terhadap distribusi data. Naïve Bayes bekerja optimal ketika data mengikuti pola distribusi tertentu, seperti distribusi normal pada data numerik. Namun, jika data yang digunakan memiliki distribusi yang tidak sesuai atau memiliki banyak variasi yang tidak terstruktur, maka performa model dapat menurun. Hal ini membuat algoritma ini kurang fleksibel dibandingkan metode lain yang mampu menangani data dengan pola yang lebih kompleks.

Selain itu, Naïve Bayes juga menghadapi masalah yang dikenal sebagai zero probability. Masalah ini terjadi ketika suatu nilai fitur tidak pernah muncul dalam data latih untuk kelas tertentu, sehingga probabilitasnya menjadi nol. Kondisi ini dapat menyebabkan keseluruhan hasil prediksi menjadi tidak akurat. Untuk mengatasi hal tersebut, biasanya digunakan teknik smoothing, seperti Laplace smoothing, namun hal ini menambah langkah tambahan dalam proses perhitungan.

Kelemahan lain yang perlu diperhatikan adalah sensitivitas terhadap kualitas dan representasi data. Jika data yang digunakan tidak lengkap, tidak seimbang, atau mengandung bias, maka hasil klasifikasi juga akan terpengaruh. Selain itu, meskipun Naïve Bayes efisien, algoritma ini kurang optimal untuk dataset yang sangat besar dan memiliki banyak fitur yang saling berkorelasi. Oleh karena itu, penggunaan Naïve Bayes perlu disesuaikan dengan karakteristik data agar hasil yang diperoleh tetap relevan dan dapat diandalkan.

3.3. Perbandingan dengan Algoritma Lain

Dalam berbagai studi komparatif, algoritma Naïve Bayes sering dibandingkan dengan metode klasifikasi lain seperti Decision Trees, k-Nearest Neighbors, Support Vector Machines, Random Forest, dan Artificial Neural Networks dalam memprediksi prestasi akademik siswa. Perbandingan ini dilakukan untuk mengetahui kelebihan dan kekurangan masing-masing algoritma dalam menangani data pendidikan yang memiliki karakteristik

tertentu. Setiap algoritma memiliki pendekatan yang berbeda dalam memproses data, sehingga hasil yang diperoleh juga dapat bervariasi tergantung pada kompleksitas dan kualitas data yang digunakan. Tahun 2014 tentang Hak Cipta

Naïve Bayes dikenal unggul dalam hal kesederhanaan dan efisiensi komputasi dibandingkan algoritma lain yang lebih kompleks seperti SVM, Random Forest, atau ANN. Decision Trees menawarkan interpretasi yang mudah dipahami, sementara KNN mengandalkan kedekatan antar data, dan ANN mampu menangani pola yang sangat kompleks. Namun, algoritma-algoritma tersebut umumnya membutuhkan waktu komputasi yang lebih besar dan sumber daya yang lebih tinggi. Dalam konteks data akademik yang relatif sederhana, Naïve Bayes sering kali mampu memberikan hasil yang kompetitif dengan waktu pemrosesan yang lebih cepat.

Dengan demikian, pemilihan algoritma dalam penelitian prediksi prestasi akademik sangat bergantung pada kebutuhan dan kondisi data yang tersedia. Naïve Bayes menjadi pilihan yang tepat ketika dibutuhkan metode yang cepat, sederhana, dan cukup akurat, terutama untuk dataset berukuran kecil hingga menengah. Sementara itu, algoritma lain dapat dipertimbangkan apabila diperlukan analisis yang lebih kompleks dan mendalam. Oleh karena itu, studi komparatif sangat penting untuk menentukan metode yang paling sesuai dalam meningkatkan kualitas prediksi di bidang pendidikan (Wu et al., 2008; Kotsiantis et al., 2007; Romero & Ventura, 2010).

Hasil dari berbagai studi komparatif menunjukkan bahwa algoritma Naïve Bayes mampu memberikan kinerja yang cukup baik dan stabil dalam berbagai kondisi, terutama pada dataset yang sederhana hingga menengah. Keunggulan utamanya terletak pada kecepatan komputasi dan kemudahan implementasi, sehingga sangat cocok digunakan pada lingkungan dengan keterbatasan sumber daya. Namun demikian, performa Naïve Bayes cenderung terbatas ketika dihadapkan pada data yang

kompleks dan memiliki hubungan antar atribut yang tidak independen.

Di sisi lain, algoritma yang lebih kompleks seperti Support Vector Machines, Random Forest, dan Artificial Neural Networks sering kali menghasilkan tingkat akurasi yang lebih tinggi. Hal ini terutama terlihat ketika dataset yang digunakan berukuran besar, memiliki banyak fitur, serta menunjukkan hubungan nonlinier antar atribut. Algoritma-algoritma tersebut mampu menangkap pola yang lebih kompleks dalam data, meskipun membutuhkan sumber daya komputasi yang lebih besar dan proses yang lebih rumit. Oleh karena itu, pemilihan algoritma harus disesuaikan dengan karakteristik data dan kebutuhan analisis agar diperoleh hasil yang optimal (Ahmed, 2024; Yağcı, 2022; Khairy et al., 2024; Aggarwal, 2015).

Namun demikian, pada kondisi tertentu, Naïve Bayes justru menunjukkan keunggulan dibandingkan algoritma lain seperti KNN atau C4.5. Beberapa penelitian melaporkan bahwa pada dataset yang relatif kecil, tereduksi, atau memiliki banyak atribut kategorikal, Naïve Bayes mampu menghasilkan nilai akurasi, precision, dan recall yang lebih baik serta lebih stabil dibandingkan algoritma berbasis jarak atau pohon keputusan (Adane et al., 2023; Yousif & Elfahal, 2023; Pandey & Pal, 2011). Hal ini disebabkan oleh kesederhanaan model Naïve Bayes dan kemampuannya dalam menghindari overfitting pada data yang terbatas (Mitchell, 1997; Han et al., 2012).

Selain pendekatan klasifikasi tunggal, beberapa penelitian juga mengembangkan variasi ensemble dari algoritma Naïve Bayes, seperti bagging dan boosting, untuk meningkatkan performa prediksi. Teknik ensemble bekerja dengan menggabungkan beberapa model dasar sehingga dapat mengurangi kesalahan prediksi dan meningkatkan kemampuan generalisasi model. Dalam konteks ini, Naïve Bayes digunakan sebagai base learner yang kemudian diperkuat melalui pendekatan ensemble untuk menghasilkan model yang lebih robust.

Hasil penelitian menunjukkan bahwa penerapan teknik ensemble pada Naïve Bayes mampu meningkatkan metrik evaluasi seperti recall, F1-score, serta stabilitas model secara keseluruhan. Peningkatan ini *зачастую* terlihat pada dataset yang tidak seimbang, di mana jumlah data pada setiap kelas tidak proporsional. Dengan adanya teknik seperti bagging dan boosting, model menjadi lebih sensitif terhadap kelas minoritas sehingga mampu memberikan prediksi yang lebih adil dan akurat dibandingkan dengan Naïve Bayes klasik (imbalanced data) (Peretz et al., 2024; Febriyanto & Wasilah, 2025; Dietterich, 2000). Temuan ini menunjukkan bahwa meskipun Naïve Bayes merupakan algoritma yang sederhana, penerapannya dapat menjadi jauh lebih kuat ketika dikombinasikan dengan teknik ensemble seperti bagging dan boosting. Kombinasi ini mampu meningkatkan kinerja model dalam hal akurasi, recall, serta stabilitas prediksi, sehingga hasil yang diperoleh menjadi lebih andal. Dalam konteks prediksi prestasi akademik siswa, pendekatan ini memberikan solusi yang efektif karena mampu mengatasi keterbatasan model tunggal sekaligus tetap mempertahankan efisiensi komputasi. Dengan demikian, integrasi Naïve Bayes dan teknik ensemble menjadi alternatif yang potensial dalam mengembangkan sistem prediksi yang lebih akurat dan terpercaya di bidang pendidikan (Aggarwal, 2015; Witten et al., 2016).



TAHAPAN MACHINE LEARNING

Untuk membangun model prediksi yang solid dalam konteks prediksi prestasi siswa, pemahaman terhadap tahapan dalam proses machine learning menjadi sangat penting. Setiap tahap dalam pipeline machine learning memiliki peran yang saling terhubung dan berkontribusi terhadap kualitas hasil akhir model. Proses yang terstruktur ini memastikan bahwa model tidak hanya mampu melakukan prediksi, tetapi juga memberikan hasil yang akurat dan dapat diandalkan dalam mendukung pengambilan keputusan di bidang pendidikan.

Tahap pertama adalah pengumpulan data, yang menjadi fondasi utama dalam pengembangan model. Data yang dikumpulkan harus relevan dan mencerminkan kondisi nyata siswa, seperti nilai akademik, kehadiran, dan aktivitas belajar. Kualitas dan kelengkapan data sangat menentukan keberhasilan model, karena data yang tidak representatif dapat menghasilkan prediksi yang bias atau kurang akurat. Oleh karena itu, proses ini harus dilakukan secara cermat dan sistematis.

Selanjutnya adalah tahap pembersihan dan preprocessing data, yang bertujuan untuk meningkatkan kualitas data sebelum digunakan dalam pelatihan model. Pada tahap ini, data yang tidak lengkap, duplikat, atau tidak konsisten akan diperbaiki atau dihapus. Selain itu, dilakukan juga transformasi data seperti normalisasi dan encoding agar data dapat diproses dengan baik oleh algoritma machine learning. Tahap ini sangat penting karena kualitas data yang baik akan membantu model dalam mempelajari pola dengan lebih efektif.

Tahap terakhir adalah pelatihan dan evaluasi model. Pada tahap pelatihan, algoritma digunakan untuk membangun model berdasarkan pola yang ditemukan dalam data. Setelah itu, model dievaluasi menggunakan berbagai metrik untuk mengukur kinerjanya. Evaluasi ini bertujuan untuk memastikan bahwa model mampu memberikan prediksi yang akurat dan dapat digunakan dalam kondisi nyata. Dengan melalui seluruh tahapan tersebut secara sistematis, model prediksi prestasi siswa yang dihasilkan akan lebih optimal dan bermanfaat bagi dunia pendidikan.

4.1. Pengumpulan Data

Langkah pertama yang krusial dalam pipeline *machine learning* adalah pengumpulan data. Dalam konteks prediksi prestasi siswa, data yang dikumpulkan meliputi berbagai informasi yang relevan dengan prestasi akademik siswa, seperti nilai mata pelajaran, kehadiran siswa, aktivitas dalam sistem manajemen pembelajaran (*Learning Management System*, LMS), serta data demografi dan perilaku siswa. Data ini dapat diperoleh dari sistem akademik sekolah, platform LMS, serta sumber data lain yang digunakan dalam proses belajar mengajar (Edmond et al., 2025; Chhimwal & Sinha, 2025). Kualitas data yang dikumpulkan sangat memengaruhi kualitas model prediksi yang dibangun. Oleh karena itu, sangat penting untuk mengumpulkan data yang lengkap dan relevan agar model yang dihasilkan dapat memberikan prediksi yang akurat.

4.2. Pembersihan dan Preprocessing Data

Setelah data terkumpul, langkah berikutnya adalah pembersihan dan preprocessing data. Data yang terkumpul sering kali mengandung nilai yang hilang (missing values), outlier, noise, dan inkonsistensi yang dapat memengaruhi kinerja model. Oleh karena itu, tahap ini bertujuan untuk menangani masalah-masalah tersebut agar data siap digunakan untuk pelatihan model.

Penanganan missing values dilakukan dengan teknik imputasi, di mana nilai yang hilang diisi dengan nilai rata-rata (mean), median, atau modus, tergantung pada jenis data dan distribusi yang diinginkan (Villegas-Ch. et al., 2023; Yasodha, 2025). Selanjutnya, data perlu dibersihkan dari outlier yang dapat mengganggu hasil analisis. Outlier yang tidak wajar atau tidak mencerminkan kinerja mayoritas siswa harus ditangani dengan hati-hati. Teknik normalisasi atau standarisasi juga diperlukan untuk memastikan bahwa fitur numerik memiliki skala yang seragam, yang penting bagi banyak algoritma machine learning untuk berfungsi dengan optimal (Nevin et al., 2025; Villegas-Ch. et al., 2023). Selain itu, untuk data kategorikal seperti jenis kelamin atau mata pelajaran, dilakukan encoding agar dapat diproses oleh model machine learning (Yasodha, 2025; Chhimwal & Sinha, 2025).

4.3. Pembagian Data dan Validasi

Setelah data siap, langkah berikutnya adalah pembagian data dan validasi. Data dibagi menjadi dua bagian utama: satu bagian untuk pelatihan model (training set) dan satu bagian untuk pengujian model (test set). Pembagian data ini umumnya dilakukan dengan rasio 75:25 atau 80:20 antara data pelatihan dan data pengujian (Kord et al., 2025).

Selain itu, untuk memastikan kestabilan dan generalisasi model, teknik k-fold cross-validation sering digunakan. Dalam teknik ini, dataset dibagi menjadi k bagian (misalnya 5 atau 10 fold), dan model dilatih dan diuji secara bergantian pada setiap bagian data.

Pendekatan ini membantu mengurangi risiko overfitting, di mana model terlalu terlatih pada data pelatihan dan tidak dapat menggeneralisasi dengan baik pada data uji yang baru (Bertolini et al., 2021; Agrawal et al., 2025). K-fold cross-validation juga berguna dalam memilih hyperparameter terbaik untuk model, yang dapat meningkatkan kinerja model prediksi.

4.4. Pelatihan Model

Setelah data dibagi dan divalidasi, langkah selanjutnya adalah pelatihan model. Pada tahap ini, berbagai algoritma supervised learning diterapkan untuk melatih model prediksi. Algoritma yang umum digunakan untuk memprediksi prestasi siswa antara lain Naïve Bayes, Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), k-Nearest Neighbors (kNN), dan deep learning (Rodríguez et al., 2025; Hussain, 2025).

Setiap algoritma memiliki kelebihan dan kekurangan tergantung pada jenis data dan tujuan analisis. Sebagai contoh, Naïve Bayes sering digunakan sebagai algoritma pembandingan karena kesederhanaannya dan kemudahan interpretasi. Algoritma lain seperti SVM atau deep learning dapat memberikan hasil yang lebih akurat pada dataset yang lebih besar atau lebih kompleks.

4.5. Evaluasi Model

Setelah model dilatih, langkah selanjutnya adalah evaluasi model untuk menilai seberapa baik model tersebut dalam memprediksi prestasi siswa. Beberapa metrik yang digunakan dalam evaluasi model antara lain akurasi, precision, recall, F1 score, dan ROC-AUC. Akurasi mengukur seberapa banyak prediksi yang benar dari seluruh prediksi yang dilakukan model, sementara precision dan recall memberikan gambaran tentang kemampuan model dalam mendeteksi siswa berisiko gagal atau berprestasi (Rodríguez et al., 2025; Srinadi et al., 2025).

Tidak untuk diperjualbelikan

F1-score adalah metrik penting yang menyeimbangkan precision dan recall, yang sangat relevan dalam konteks prediksi prestasi siswa, karena sering kali kita ingin menghindari kesalahan dalam mendeteksi siswa yang berisiko gagal (Deshpande et al., 2025). Evaluasi ini memberikan gambaran yang jelas mengenai kinerja model dan memungkinkan pemilihan model terbaik.

4.6. Optimasi dan Deployment

Langkah terakhir adalah optimasi dan deployment model. Setelah model dievaluasi, tahap optimasi dilakukan untuk meningkatkan kinerja model, seperti melakukan hyperparameter tuning dengan menggunakan teknik grid search atau optimasi berbasis metaheuristik (Aldhahri et al., 2025; Rodríguez et al., 2025). Setelah mendapatkan model yang optimal, model tersebut siap untuk deployment, yaitu diintegrasikan ke dalam sistem pendidikan, seperti early warning system, yang memberi rekomendasi kepada pendidik mengenai siswa yang berisiko gagal atau membutuhkan bantuan lebih lanjut.

Model yang telah dioptimasi juga dapat digunakan untuk memberikan rekomendasi mata pelajaran atau membantu dalam perencanaan akademik siswa. Dengan demikian, model ini dapat memberikan kontribusi nyata dalam meningkatkan kualitas pendidikan dan memastikan bahwa setiap siswa mendapatkan perhatian yang sesuai dengan kebutuhannya (EdmondEdmond et al., 2025; Chhimwal & Sinha, 2025).



Tidak untuk diperjualbelikan

Dilindungi oleh Undang-Undang Republik Indonesia
Nomor 28 Tahun 2014 tentang Hak Cipta

©p.pustakariyadz



Tidak untuk diperjualbelikan

EVALUASI MODEL



Evaluasi model merupakan tahapan yang sangat penting dalam proses machine learning, khususnya dalam permasalahan klasifikasi. Tahap ini bertujuan untuk mengukur seberapa baik algoritma yang digunakan mampu memprediksi atau mengelompokkan data sesuai dengan kondisi sebenarnya. Tanpa adanya evaluasi, sulit untuk mengetahui apakah model yang dibangun sudah layak digunakan atau masih memerlukan perbaikan. Oleh karena itu, evaluasi menjadi langkah kunci dalam memastikan kualitas dan keandalan model.

Dalam proses evaluasi, model diuji menggunakan data tertentu yang tidak digunakan saat pelatihan, sehingga dapat memberikan gambaran objektif mengenai performa model. Berbagai metrik seperti akurasi, precision, recall, dan F1-score digunakan untuk menilai kinerja model dari berbagai sudut pandang. Masing-masing metrik memiliki fungsi yang berbeda-beda, sehingga penggunaannya secara bersama-sama dapat memberikan evaluasi yang lebih komprehensif terhadap kemampuan model dalam menangani data.

Tidak untuk diperjualbelikan

Dengan demikian, evaluasi model tidak hanya berfungsi sebagai alat ukur, tetapi juga sebagai dasar untuk melakukan perbaikan dan pengembangan model lebih lanjut. Hasil evaluasi dapat digunakan untuk mengidentifikasi kelemahan model, seperti kesalahan klasifikasi atau ketidakseimbangan data, sehingga dapat dilakukan penyesuaian yang diperlukan. Melalui evaluasi yang tepat, model machine learning yang dihasilkan akan lebih akurat, efisien, dan dapat diandalkan dalam mendukung pengambilan keputusan (Han et al., 2012; Witten et al., 2016).

Dalam konteks prediksi prestasi akademik siswa, evaluasi model memungkinkan peneliti untuk mengukur kemampuan model dalam memprediksi nilai mata pelajaran atau kelulusan siswa secara akurat dan konsisten (Romero & Ventura, 2010; Yağcı, 2022). Pemilihan teknik evaluasi yang tepat merupakan faktor penting dalam memastikan bahwa model machine learning yang dikembangkan memiliki kinerja yang optimal dan dapat dipercaya. Setiap teknik evaluasi memiliki karakteristik dan tujuan yang berbeda, sehingga harus disesuaikan dengan jenis data dan permasalahan yang dihadapi. Dalam konteks klasifikasi, penggunaan metrik seperti akurasi saja tidak selalu cukup, terutama jika data yang digunakan tidak seimbang. Oleh karena itu, diperlukan kombinasi metrik evaluasi agar hasil penilaian model lebih komprehensif.

Dalam bidang pendidikan, ketepatan dalam mengevaluasi model menjadi sangat krusial karena hasilnya akan berdampak langsung pada pengambilan keputusan akademik. Misalnya, dalam menentukan siswa berprestasi atau siswa yang membutuhkan perhatian khusus, kesalahan dalam evaluasi model dapat menyebabkan keputusan yang kurang tepat. Oleh karena itu, teknik evaluasi seperti precision, recall, F1-score, serta validasi silang (cross-validation) perlu digunakan untuk memastikan bahwa model benar-benar mampu bekerja dengan baik dalam berbagai kondisi data.

Dengan demikian, pemilihan teknik evaluasi yang tepat tidak hanya meningkatkan akurasi model, tetapi juga menjamin keandalan hasil yang dihasilkan. Model yang telah dievaluasi dengan metode yang sesuai akan lebih mampu memberikan prediksi yang konsisten dan dapat dipertanggungjawabkan. Hal ini sangat penting dalam dunia pendidikan, di mana setiap keputusan yang diambil harus didasarkan pada analisis yang objektif, akurat, dan mendukung peningkatan kualitas pembelajaran (Kotsiantis et al., 2007).

Metrik evaluasi yang umum digunakan dalam klasifikasi meliputi akurasi, precision, recall, dan F1-score (Powers, 2011; Aggarwal, 2015). Akurasi mengukur proporsi prediksi yang benar terhadap keseluruhan data, namun metrik ini dapat menjadi kurang representatif pada dataset yang tidak seimbang (imbalanced data) (Sokolova & Lapalme, 2009). Oleh karena itu, precision dan recall digunakan untuk memberikan gambaran yang lebih rinci mengenai kemampuan model dalam mengklasifikasikan data secara benar, khususnya pada kelas tertentu, seperti siswa yang lulus atau tidak lulus (Chawla et al., 2004; Yağcı, 2022).

F1-score merupakan rata-rata harmonik dari precision dan recall, yang digunakan untuk menilai keseimbangan antara kedua metrik tersebut dan sering dijadikan indikator utama dalam evaluasi model klasifikasi pendidikan (Sokolova & Lapalme, 2009; Khairy et al., 2024). Dalam penelitian prediksi prestasi siswa, metrik-metrik ini digunakan untuk mengevaluasi kinerja algoritma Naïve Bayes dan membandingkannya dengan algoritma klasifikasi lain guna menentukan model yang paling efektif dan stabil (Romero & Ventura, 2010; Ahmed, 2024; Witten et al., 2016).

5.1. Metrik Evaluasi

Evaluasi model machine learning dalam tugas klasifikasi sering dilakukan menggunakan confusion matrix, yaitu suatu tabel yang digunakan untuk menggambarkan kinerja model dalam mengklasifikasikan data. Confusion matrix memberikan informasi

yang lebih rinci dibandingkan hanya melihat nilai akurasi, karena mampu menunjukkan jumlah prediksi yang benar maupun salah untuk setiap kelas. Dengan demikian, metode ini menjadi alat yang penting untuk memahami performa model secara lebih mendalam.

Confusion matrix terdiri dari empat komponen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). True Positive adalah jumlah data yang benar diklasifikasikan sebagai kelas positif, sedangkan True Negative adalah jumlah data yang benar diklasifikasikan sebagai kelas negatif. False Positive terjadi ketika data negatif salah diklasifikasikan sebagai positif, sementara False Negative terjadi ketika data positif salah diklasifikasikan sebagai negatif. Keempat komponen ini menjadi dasar dalam menghitung berbagai metrik evaluasi seperti akurasi, precision, recall, dan F1-score.

Dengan menggunakan confusion matrix, peneliti atau pengguna dapat mengidentifikasi jenis kesalahan yang paling sering terjadi pada model. Hal ini sangat penting untuk melakukan perbaikan model agar hasil prediksi menjadi lebih akurat dan seimbang. Dalam konteks pendidikan, pemahaman terhadap confusion matrix membantu memastikan bahwa model klasifikasi, seperti dalam penentuan siswa berprestasi, dapat bekerja secara optimal dan meminimalkan kesalahan yang berdampak pada pengambilan keputusan akademik (Sokolova & Lapalme, 2009; Powers, 2011). Confusion matrix memberikan gambaran rinci mengenai kesalahan dan keberhasilan model dalam mengklasifikasikan data, sehingga menjadi dasar dalam perhitungan berbagai metrik evaluasi utama seperti akurasi, precision, recall, dan F1-score (Han et al., 2012; Witten et al., 2016).

Akurasi merupakan salah satu metrik evaluasi yang paling umum digunakan dalam menilai kinerja model klasifikasi. Menurut Aggarwal (2015), akurasi mengukur proporsi prediksi yang benar dibandingkan dengan seluruh prediksi yang dilakukan oleh model. Metrik ini memberikan gambaran sederhana mengenai seberapa

sering model berhasil mengklasifikasikan data dengan tepat, sehingga sering dijadikan indikator awal dalam mengevaluasi performa suatu model.

Secara matematis, akurasi dihitung dengan membandingkan jumlah prediksi benar, yaitu True Positive (TP) dan True Negative (TN), dengan total keseluruhan prediksi yang meliputi TP, TN, False Positive (FP), dan False Negative (FN). Dengan kata lain, akurasi menunjukkan persentase keberhasilan model dalam mengklasifikasikan data secara keseluruhan. Nilai akurasi yang tinggi biasanya menunjukkan bahwa model memiliki performa yang baik dalam mengenali pola data.

Namun demikian, penggunaan akurasi sebagai satu-satunya metrik evaluasi memiliki keterbatasan, terutama pada dataset yang tidak seimbang. Dalam kondisi tersebut, model dapat menghasilkan nilai akurasi yang tinggi tetapi sebenarnya kurang efektif dalam mengklasifikasikan kelas minoritas. Oleh karena itu, meskipun akurasi penting, penggunaannya sebaiknya dilengkapi dengan metrik lain seperti precision, recall, dan F1-score agar evaluasi model menjadi lebih komprehensif dan akurat (Powers, 2011). Meskipun akurasi memberikan gambaran umum tentang kinerja model, metrik ini dapat bersifat menyesatkan ketika dataset memiliki distribusi kelas yang tidak seimbang (imbalanced dataset), seperti pada kasus prediksi siswa berisiko gagal yang jumlahnya relatif lebih sedikit dibandingkan siswa yang lulus (Sokolova & Lapalme, 2009; Chawla et al., 2004).

Precision merupakan salah satu metrik evaluasi dalam klasifikasi yang digunakan untuk mengukur tingkat ketepatan prediksi positif yang dihasilkan oleh model. Metrik ini menunjukkan seberapa banyak data yang diprediksi sebagai kelas positif, seperti siswa “berisiko” atau “gagal”, benar-benar sesuai dengan kondisi sebenarnya. Dengan kata lain, precision berfokus pada kualitas dari prediksi positif yang dihasilkan oleh model, bukan hanya jumlahnya.

Dalam konteks pendidikan, precision sangat penting ketika model digunakan untuk mengidentifikasi siswa yang membutuhkan perhatian khusus. Jika nilai precision tinggi, maka sebagian besar siswa yang diklasifikasikan sebagai “berisiko” memang benar-benar berada dalam kondisi tersebut. Hal ini membantu pihak sekolah untuk memberikan intervensi yang tepat sasaran tanpa membuang sumber daya pada siswa yang sebenarnya tidak memerlukan bantuan tambahan.

Namun, precision juga memiliki keterbatasan jika digunakan secara tunggal. Model dengan precision tinggi belum tentu mampu mendeteksi seluruh siswa yang benar-benar berisiko, karena mungkin ada sebagian yang tidak teridentifikasi. Oleh karena itu, penggunaan precision sebaiknya dikombinasikan dengan metrik lain seperti recall agar evaluasi model menjadi lebih seimbang dan memberikan gambaran yang lebih lengkap terhadap kinerja sistem klasifikasi (Powers, 2011; Han et al., 2012). Precision dihitung sebagai perbandingan antara True Positive (TP) dan total prediksi positif (TP + FP). Dalam konteks pendidikan, precision sangat penting karena kesalahan dalam mengklasifikasikan siswa sebagai berisiko dapat menyebabkan intervensi yang tidak perlu, pemborosan sumber daya, atau kebijakan akademik yang kurang tepat (Romero & Ventura, 2010; Yağcı, 2022).

Recall, yang juga dikenal sebagai sensitivitas, mengukur kemampuan model dalam mendeteksi seluruh siswa yang benar-benar berisiko gagal (Sokolova & Lapalme, 2009). Recall dihitung dengan membandingkan jumlah True Positive (TP) dengan total kasus positif sebenarnya (TP + FN). Metrik ini sangat penting dalam sistem peringatan dini (early warning system) di bidang pendidikan, karena tujuan utama sistem tersebut adalah meminimalkan jumlah siswa berisiko yang tidak terdeteksi oleh model (Baker & Yacef, 2009; Márquez-Vera et al., 2016).

F1-score merupakan rata-rata harmonik dari precision dan recall, yang digunakan untuk memberikan evaluasi yang seimbang

terhadap kinerja model, terutama ketika terdapat trade-off antara kedua metrik tersebut (Powers, 2011; Witten et al., 2016). F1-score merupakan metrik evaluasi yang sangat berguna dalam kasus prediksi prestasi akademik siswa karena mampu memberikan keseimbangan antara precision dan recall. Dalam konteks ini, model tidak hanya diharapkan tepat dalam mengidentifikasi siswa yang berisiko, tetapi juga mampu menangkap sebanyak mungkin siswa yang действительно membutuhkan perhatian. F1-score membantu memastikan bahwa model tidak hanya fokus pada satu aspek saja, melainkan mempertimbangkan ketepatan dan kelengkapan hasil prediksi secara bersamaan.

Penggunaan F1-score menjadi sangat penting karena kesalahan klasifikasi dapat berdampak langsung pada pengambilan keputusan pendidikan. Jika model menghasilkan terlalu banyak false positive, maka siswa yang sebenarnya tidak berisiko bisa salah dikategorikan dan mendapatkan perlakuan yang tidak sesuai. Sebaliknya, jika terlalu banyak false negative, siswa yang benar-benar membutuhkan bantuan justru terabaikan. Oleh karena itu, F1-score memberikan ukuran yang lebih seimbang untuk memastikan bahwa model bekerja secara optimal dan mendukung keputusan yang lebih akurat serta adil dalam lingkungan pendidikan (Romero & Ventura, 2010; Ahmed, 2024).

5.2. Cross Validation

Salah satu teknik yang paling umum digunakan untuk menilai kestabilan dan kemampuan generalisasi model machine learning adalah k-fold cross-validation (Kohavi, 1995; Witten et al., 2016). Dalam teknik ini, dataset dibagi menjadi k bagian dengan ukuran yang relatif sama. Model kemudian dilatih sebanyak k kali, di mana pada setiap iterasi satu bagian digunakan sebagai data uji (testing set), sedangkan k-1 bagian lainnya digunakan sebagai data pelatihan (training set) (Hastie et al., 2009; Han et al., 2012). Hasil evaluasi dari setiap iterasi kemudian dirata-ratakan untuk menghasilkan estimasi kinerja model yang lebih stabil dan andal,

menggunakan metrik seperti akurasi, precision, recall, dan F1-score (Powers, 2011; Sokolova & Lapalme, 2009).

Penggunaan k-fold cross-validation merupakan salah satu teknik evaluasi yang banyak digunakan dalam machine learning untuk meningkatkan keandalan hasil pengujian model. Salah satu keunggulan utama metode ini adalah kemampuannya dalam mengurangi risiko overfitting, yaitu kondisi ketika model terlalu menyesuaikan diri dengan data pelatihan sehingga tidak mampu bekerja dengan baik pada data baru (Hastie et al., 2009; Aggarwal, 2015). Dengan membagi data ke dalam beberapa lipatan (fold), model dilatih dan diuji secara bergantian sehingga tidak hanya bergantung pada satu subset data tertentu.

Dalam k-fold cross-validation, seluruh data digunakan secara bergiliran sebagai data latih dan data uji. Hal ini memungkinkan model untuk diuji pada berbagai variasi data, sehingga hasil evaluasi menjadi lebih stabil dan representatif. Dibandingkan dengan metode hold-out yang hanya menggunakan satu kali pembagian data, cross-validation mampu memberikan gambaran performa model yang lebih akurat karena meminimalkan bias akibat pemilihan data tertentu (Kohavi, 1995; Witten et al., 2016).

Dengan demikian, k-fold cross-validation menjadi teknik yang efektif dalam mengevaluasi model secara menyeluruh. Pendekatan ini tidak hanya membantu meningkatkan keakuratan penilaian performa model, tetapi juga memberikan keyakinan bahwa model yang dihasilkan dapat bekerja dengan baik pada data baru. Oleh karena itu, metode ini sangat direkomendasikan dalam pengembangan model machine learning, termasuk dalam konteks pendidikan untuk memastikan hasil prediksi yang lebih andal dan dapat dipertanggungjawabkan.

Selain itu, cross-validation sering digunakan dalam proses pemilihan hyperparameter melalui teknik seperti grid search atau random search, di mana berbagai kombinasi parameter diuji untuk memperoleh konfigurasi model dengan kinerja terbaik (Bergstra & Bengio, 2012; Aggarwal, 2015). Pendekatan ini sangat

penting dalam meningkatkan performa algoritma klasifikasi, termasuk Naïve Bayes, terutama ketika digunakan pada dataset pendidikan yang memiliki variasi karakteristik siswa (Romero & Ventura, 2010; Yağcı, 2022).

Dalam konteks Educational Data Mining (EDM), penggunaan k-fold cross-validation sering dikombinasikan dengan teknik feature selection untuk meningkatkan kualitas model prediksi prestasi akademik siswa. Kombinasi ini memungkinkan proses evaluasi model menjadi lebih akurat sekaligus memastikan bahwa model tidak hanya bergantung pada seluruh atribut yang tersedia, tetapi pada fitur-fitur yang benar-benar relevan. Dengan demikian, model yang dihasilkan menjadi lebih efisien dan memiliki kemampuan generalisasi yang lebih baik terhadap data baru (Bertolini et al., 2021; Hoyos & Caicedo-Castro, 2024).

Feature selection berperan penting dalam memilih atribut yang paling berpengaruh terhadap hasil prediksi, seperti nilai mata pelajaran, tingkat kehadiran, dan indikator keterlibatan siswa dalam pembelajaran. Dengan mengurangi jumlah fitur yang tidak relevan atau redundan, kompleksitas model dapat ditekan sehingga proses komputasi menjadi lebih cepat dan hasil prediksi lebih akurat. Selain itu, model yang lebih sederhana cenderung lebih stabil dan tidak mudah mengalami overfitting (Guyon & Elisseeff, 2003).

Lebih dari itu, penerapan feature selection juga memberikan manfaat dari sisi interpretabilitas model. Dengan hanya menggunakan fitur-fitur penting, pendidik dapat lebih mudah memahami faktor-faktor yang memengaruhi prestasi siswa dan mengambil keputusan yang tepat berdasarkan hasil analisis tersebut. Hal ini sejalan dengan tujuan EDM yang tidak hanya menghasilkan prediksi yang akurat, tetapi juga memberikan wawasan yang dapat digunakan untuk meningkatkan kualitas pembelajaran dan pengelolaan pendidikan (Romero & Ventura, 2010).

5.3. Tantangan Evaluasi

Dalam konteks prediksi prestasi akademik siswa, penafsiran metrik evaluasi model machine learning menghadapi sejumlah tantangan yang cukup kompleks. Salah satu tantangan utama adalah ketidakseimbangan kelas (class imbalance), yaitu kondisi di mana jumlah siswa yang tergolong gagal atau berisiko jauh lebih sedikit dibandingkan dengan siswa yang lulus atau berprestasi baik (He & Garcia, 2009; Johnson & Khoshgoftaar, 2019). Ketidakseimbangan ini dapat memengaruhi kemampuan model dalam mengenali pola dari kelas minoritas, sehingga hasil prediksi menjadi kurang optimal.

Pada kondisi tersebut, penggunaan metrik akurasi sering kali menjadi menyesatkan. Model dapat menghasilkan nilai akurasi yang tinggi hanya dengan memprediksi sebagian besar data sebagai kelas mayoritas, tanpa benar-benar mampu mengidentifikasi siswa yang berisiko. Hal ini tentu menjadi masalah serius dalam konteks pendidikan, karena siswa yang membutuhkan perhatian justru tidak terdeteksi dengan baik (Adnan et al., 2022; Chhimwal & Sinha, 2025).

Oleh karena itu, diperlukan pendekatan evaluasi yang lebih komprehensif dengan mempertimbangkan metrik lain seperti precision, recall, dan F1-score. Metrik-metrik ini memberikan gambaran yang lebih jelas mengenai kemampuan model dalam mendeteksi kelas minoritas. Dengan pemilihan metrik evaluasi yang tepat, model prediksi dapat digunakan secara lebih efektif untuk mendukung pengambilan keputusan yang adil dan akurat dalam dunia pendidikan, khususnya dalam mengidentifikasi siswa yang benar-benar membutuhkan perhatian khusus.

Oleh karena itu, metrik evaluasi seperti precision, recall, dan F1-score menjadi lebih relevan untuk digunakan dalam menilai kinerja model prediksi prestasi siswa (He & Garcia, 2009; Branco et al., 2016). Recall sangat penting dalam konteks pendidikan karena berkaitan langsung dengan kemampuan model dalam

mendeteksi siswa yang berisiko gagal, sedangkan precision membantu memastikan bahwa intervensi akademik tidak diberikan secara berlebihan kepada siswa yang sebenarnya tidak membutuhkan (Adnan et al., 2022; Sun et al., 2023). Dengan demikian, F1-score sering digunakan sebagai metrik utama karena memberikan keseimbangan antara precision dan recall pada dataset yang tidak seimbang (Branco et al., 2016).

Tantangan lain dalam evaluasi model adalah overfitting, yaitu kondisi ketika model terlalu menyesuaikan diri dengan data pelatihan sehingga kehilangan kemampuan untuk menggeneralisasi pada data baru (Goodfellow et al., 2016; Zhang et al., 2021). Overfitting biasanya ditandai dengan perbedaan kinerja yang signifikan antara data pelatihan dan data uji, di mana akurasi pelatihan sangat tinggi tetapi performa pada data uji menurun drastis (Domingos, 2012). Untuk mengatasi masalah ini, teknik evaluasi seperti cross-validation sangat penting digunakan agar model diuji pada berbagai subset data dan menghasilkan estimasi kinerja yang lebih stabil dan representatif (Rimal & Sharma, 2025; Zhang et al., 2021).

Selain aspek kinerja, interpretabilitas model menjadi tantangan penting dalam penerapan machine learning di bidang pendidikan. Meskipun metrik evaluasi seperti akurasi dan F1-score dapat menunjukkan tingkat keberhasilan model dalam melakukan prediksi, metrik tersebut tidak mampu menjelaskan alasan di balik hasil yang dihasilkan (Doshi-Velez & Kim, 2017; Molnar, 2022). Hal ini menjadi keterbatasan karena model yang hanya memberikan hasil tanpa penjelasan sulit untuk dipahami secara menyeluruh oleh pengguna.

Dalam praktik pendidikan, interpretabilitas sangat dibutuhkan oleh pendidik dan pengambil kebijakan untuk memahami faktor-faktor yang memengaruhi hasil prediksi. Mereka tidak hanya membutuhkan model yang akurat, tetapi juga model yang mampu memberikan penjelasan yang jelas dan logis. Dengan adanya interpretasi yang baik, hasil prediksi dapat digunakan sebagai

dasar untuk merancang intervensi yang tepat, seperti memberikan bimbingan tambahan kepada siswa yang berisiko atau menyesuaikan strategi pembelajaran (Bertolini et al., 2021; Lundberg et al., 2020).

Model yang sulit diinterpretasikan, meskipun memiliki kinerja tinggi, berpotensi menimbulkan keraguan dan menurunkan kepercayaan pengguna. Hal ini dapat menghambat penerapan teknologi tersebut dalam sistem pendidikan nyata. Oleh karena itu, penting untuk mengembangkan atau memilih model yang tidak hanya memiliki performa yang baik, tetapi juga mudah dipahami. Dengan demikian, hasil prediksi dapat dimanfaatkan secara optimal dan bertanggung jawab dalam mendukung pengambilan keputusan di bidang pendidikan.

5.4. Perbandingan Kinerja Model

Dalam banyak studi prediksi prestasi akademik siswa, algoritma Naïve Bayes sering dibandingkan dengan berbagai model klasifikasi lain seperti Decision Trees, Random Forest, Support Vector Machines, Artificial Neural Networks, serta metode ensemble modern seperti XGBoost (Bhardwaj & Pal, 2019; Al-Barrak & Al-Razgan, 2016; Adnan et al., 2022; Zhan & Shen, 2025). Perbandingan ini dilakukan untuk menguji kinerja masing-masing algoritma dalam menangani data pendidikan yang memiliki karakteristik unik, seperti ukuran dataset yang beragam, distribusi data yang tidak seimbang, serta keterkaitan antar variabel yang kompleks.

Tujuan utama dari studi komparatif tersebut adalah untuk menilai apakah Naïve Bayes cukup efektif digunakan sebagai model utama atau lebih tepat dijadikan sebagai baseline dalam analisis. Sebagai algoritma yang sederhana dan efisien, Naïve Bayes sering digunakan sebagai titik awal untuk membandingkan performa model lain yang lebih kompleks. Hal ini penting karena hasil perbandingan dapat menunjukkan sejauh mana peningkatan akurasi atau performa yang diberikan oleh algoritma yang lebih

canggih dibandingkan dengan metode dasar (Baker & Inventado, 2014; Shahiria et al., 2015).

Dengan demikian, studi komparatif memberikan wawasan yang lebih komprehensif mengenai kelebihan dan keterbatasan masing-masing algoritma dalam konteks pendidikan. Naïve Bayes tetap memiliki nilai penting karena kemudahan implementasi dan efisiensinya, terutama pada dataset sederhana. Namun, pada kondisi data yang lebih kompleks, algoritma lain mungkin lebih unggul. Oleh karena itu, pemilihan model yang tepat harus mempertimbangkan kebutuhan analisis, karakteristik data, serta tujuan penggunaan model dalam mendukung pengambilan keputusan pendidikan.

Beberapa penelitian menunjukkan bahwa algoritma berbasis ensemble dan margin, seperti XGBoost dan Support Vector Machines (SVM), cenderung menghasilkan kinerja yang lebih tinggi dibandingkan algoritma sederhana. Hal ini terlihat dari hasil penelitian Sihare (2024), di mana model XGBoost mampu mencapai akurasi sebesar 97%, precision 98%, recall 98%, dan F1-score 98%. Sebaliknya, algoritma Decision Tree hanya menghasilkan akurasi 75%, precision 47%, recall 72%, dan F1-score 57%. Perbedaan ini menunjukkan bahwa metode yang lebih kompleks memiliki kemampuan yang lebih baik dalam mengolah data dan menghasilkan prediksi yang lebih akurat.

Keunggulan tersebut didukung oleh penelitian lain yang menyatakan bahwa SVM dan XGBoost sangat efektif dalam menangani data yang kompleks dan memiliki hubungan nonlinier antar variabel (Cortes & Vapnik, 1995; Chen & Guestrin, 2016). Algoritma ini mampu memodelkan pola yang sulit ditangkap oleh metode sederhana, terutama pada dataset berskala besar dengan banyak fitur. Oleh karena itu, dalam konteks prediksi prestasi akademik siswa, penggunaan algoritma yang lebih canggih dapat memberikan hasil yang lebih optimal, meskipun memerlukan sumber daya komputasi yang lebih tinggi dibandingkan metode yang lebih sederhana.

Meskipun demikian, berbagai studi juga menegaskan bahwa Naïve Bayes tetap memiliki keunggulan dalam kondisi tertentu, khususnya pada dataset berukuran kecil, data berdimensi rendah, atau ketika kebutuhan utama adalah efisiensi komputasi dan interpretabilitas model (Rish, 2001; Ng & Jordan, 2002; Al-Barrak & Al-Razgan, 2016). Dalam konteks pendidikan, interpretabilitas menjadi aspek yang sangat penting karena pendidik dan pengambil kebijakan memerlukan pemahaman yang jelas mengenai faktor-faktor yang memengaruhi prediksi prestasi siswa (Baker & Inventado, 2014; Bhardwaj & Pal, 2019).

Oleh karena itu, Naïve Bayes sering digunakan sebagai algoritma pembanding dalam berbagai studi komparatif untuk mengevaluasi efektivitas model yang lebih kompleks. Sebagai metode yang sederhana dan mudah diimplementasikan, Naïve Bayes berfungsi sebagai baseline yang membantu peneliti memahami sejauh mana peningkatan performa yang dihasilkan oleh algoritma lain seperti XGBoost atau Artificial Neural Networks. Dengan adanya pembanding ini, analisis menjadi lebih objektif dan terukur dalam menilai keunggulan masing-masing model.

Meskipun tingkat akurasi Naïve Bayes sering kali lebih rendah dibandingkan algoritma yang lebih canggih, keunggulannya terletak pada kesederhanaan dan transparansi dalam proses perhitungan. Model ini mudah dipahami karena berbasis probabilitas yang jelas, sehingga hasil prediksi dapat dijelaskan dengan lebih logis. Hal ini menjadi nilai tambah, terutama dalam bidang pendidikan, di mana interpretasi hasil sangat penting untuk mendukung pengambilan keputusan yang tepat (Shahiria et al., 2015; Adnan et al., 2022; Zhan & Shen, 2025).

Dengan demikian, Naïve Bayes tetap relevan dalam penelitian prediksi prestasi siswa, khususnya ketika tujuan penelitian tidak hanya berfokus pada kinerja prediksi, tetapi juga pada kemudahan interpretasi dan penerapan praktis. Dalam lingkungan pendidikan, model yang sederhana namun dapat

dipahami sering kali lebih mudah diterima oleh pendidik dan pengambil kebijakan. Oleh karena itu, Naïve Bayes menjadi pilihan yang tepat sebagai alat analisis sekaligus pembanding dalam pengembangan sistem prediksi berbasis data..

©p.pustakariyadz

Dilindungi oleh Undang-Undang Republik Indonesia
Nomor 28 Tahun 2014 tentang Hak Cipta

©p.pustakariyadz

ALAT BANTU PEMROGRAMAN



Pemilihan alat pemrograman yang tepat merupakan faktor penting dalam pengembangan model prediksi, khususnya dalam konteks prediksi prestasi siswa menggunakan machine learning. Bahasa pemrograman yang digunakan akan sangat memengaruhi kemudahan dalam pengolahan data, implementasi algoritma, serta proses evaluasi model. Dengan memilih alat yang sesuai, pengembang dapat bekerja lebih efisien dalam mengelola data akademik yang kompleks serta membangun model yang akurat dan dapat diandalkan.

Selain itu, platform pemrograman yang tepat juga mendukung proses pengembangan yang lebih sistematis dan terstruktur. Fitur-fitur yang tersedia, seperti pustaka analisis data dan machine learning, dapat mempercepat proses pengolahan serta meningkatkan kualitas hasil prediksi. Oleh karena itu, pemilihan alat pemrograman tidak hanya berpengaruh pada kemudahan teknis, tetapi juga berkontribusi terhadap keberhasilan implementasi model dalam mendukung pengambilan keputusan di bidang pendidikan..

Di dunia machine learning, Python telah menjadi bahasa yang dominan karena kemudahan penggunaannya serta dukungan ekosistem yang sangat luas. Python memiliki sintaks yang sederhana dan mudah dipahami, sehingga cocok digunakan oleh pemula maupun profesional. Selain itu, komunitas pengguna Python yang besar dan aktif memungkinkan pengembang untuk dengan mudah menemukan referensi, dokumentasi, serta solusi terhadap berbagai permasalahan yang dihadapi (Muqri et al., 2024; Naser, 2025).

Keunggulan Python juga didukung oleh berbagai pustaka open source yang sangat membantu dalam pengolahan dan analisis data, seperti Pandas untuk manipulasi data, NumPy untuk komputasi numerik, serta Scikit-learn untuk implementasi algoritma machine learning. Dengan adanya tools tersebut, pengembangan model prediksi menjadi lebih cepat, akurat, dan efisien. Oleh karena itu, Python menjadi pilihan yang tepat dalam membangun sistem prediksi prestasi siswa yang berbasis data dan teknologi.

6.1. Penggunaan Python dalam Machine Learning

Python adalah pilihan utama dalam pengembangan algoritma machine learning karena kesederhanaannya dan kemampuannya untuk menangani berbagai jenis data. Keunggulan utama Python terletak pada kemudahan implementasi serta fleksibilitas dalam mengintegrasikan berbagai pustaka machine learning yang sudah sangat teruji. Ekosistem Python yang kaya memungkinkan para peneliti dan praktisi untuk membangun model prediksi dengan cepat dan efisien (Nikandrov, 2023; Kanungo, 2023). Dalam konteks pendidikan dan analisis data akademik, Python memungkinkan penanganan data besar dengan mudah dan memungkinkan eksperimen yang lebih cepat serta visualisasi yang lebih jelas.

Pandas merupakan salah satu pustaka yang sangat penting dalam bahasa pemrograman Python, khususnya untuk keperluan data

preprocessing dan manipulasi data. Pustaka ini menyediakan struktur data yang fleksibel seperti DataFrame dan Series, yang memungkinkan pengguna untuk mengelola data secara terstruktur dan efisien. Dalam konteks machine learning, Pandas sering digunakan pada tahap awal untuk mempersiapkan data sebelum dilakukan analisis lebih lanjut.

Dengan Pandas, berbagai operasi pengolahan data seperti pemfilteran, agregasi, dan transformasi fitur dapat dilakukan dengan mudah dan cepat. Kemampuan ini sangat berguna dalam menangani dataset berukuran besar yang umum ditemukan dalam bidang Educational Data Mining (EDM) (Noviandy et al., 2025; Naser, 2025). Pengguna dapat membersihkan data, menangani nilai yang hilang, serta mengubah format data sesuai kebutuhan algoritma machine learning dengan lebih praktis.

Selain itu, Pandas juga memudahkan pengolahan data akademik seperti nilai mata pelajaran, kehadiran siswa, dan atribut lainnya yang relevan. Data-data tersebut dapat diorganisasi, dianalisis, dan diolah menjadi bentuk yang siap digunakan untuk proses pemodelan. Dengan demikian, Pandas menjadi alat yang sangat penting dalam mendukung analisis data pendidikan secara efisien, serta membantu menghasilkan model prediksi yang lebih akurat dan terstruktur.

6.2. Jupyter Notebook

Selain alat pemrograman dan pustaka yang telah disebutkan, Jupyter Notebook merupakan platform yang sangat populer di kalangan peneliti dan praktisi machine learning. Jupyter memungkinkan eksperimen interaktif, di mana pengguna dapat menulis kode, menjalankan eksperimen, dan melihat hasil secara langsung dalam satu lingkungan yang terintegrasi (Mujilawati et al., 2021). Dengan menggunakan Jupyter Notebook, penelitian dapat dilakukan secara iteratif, memungkinkan eksperimen yang cepat dan perubahan kode tanpa perlu keluar dari lingkungan yang sama.

Jupyter Notebook merupakan salah satu alat yang sangat bermanfaat dalam pengembangan dan penelitian berbasis machine learning, khususnya dalam hal pendokumentasian eksperimen. Dengan format berbasis sel, Jupyter memungkinkan peneliti untuk menggabungkan kode program, hasil eksekusi, serta penjelasan dalam bentuk teks secara terstruktur dalam satu dokumen. Hal ini membuat proses analisis menjadi lebih terorganisir dan mudah diikuti, baik oleh peneliti sendiri maupun oleh pihak lain yang ingin memahami alur penelitian.

Selain itu, Jupyter Notebook juga sangat mendukung proses visualisasi data dan hasil model. Dengan memanfaatkan pustaka seperti Matplotlib dan Seaborn, peneliti dapat menampilkan grafik, diagram, dan visualisasi lainnya secara langsung setelah proses komputasi dilakukan (Noviandy et al., 2025). Visualisasi ini sangat penting untuk memahami pola data, mengevaluasi performa model, serta menyajikan hasil analisis secara lebih intuitif dan informatif.

Keunggulan lain dari Jupyter adalah kemampuannya dalam mendukung analisis yang transparan dan interaktif. Setiap langkah dalam proses eksperimen dapat ditampilkan secara jelas, mulai dari preprocessing data, pelatihan model, hingga evaluasi hasil. Hal ini memungkinkan proses penelitian menjadi lebih terbuka dan mudah diverifikasi, sehingga meningkatkan kepercayaan terhadap hasil yang diperoleh.

Dengan demikian, penggunaan Jupyter Notebook dalam penelitian tidak hanya mempermudah proses analisis, tetapi juga mendukung dokumentasi yang sistematis dan replikasi eksperimen di masa depan. Peneliti lain dapat dengan mudah mengikuti langkah-langkah yang telah dilakukan dan menguji kembali model yang dikembangkan. Hal ini menjadikan Jupyter sebagai alat yang sangat efektif dalam mendukung penelitian ilmiah yang transparan, akurat, dan dapat dipertanggungjawabkan.



PENERAPAN ALGORITMA NAÏVE BAYES

7.1. Desain dan Pendekatan Kasus

Studi kasus dalam konteks pengolahan data pendidikan umumnya menggunakan pendekatan kuantitatif yang menekankan pada pengukuran, analisis numerik, serta pengujian model secara sistematis. Pendekatan ini sering dipadukan dengan metode eksperimen untuk membangun dan menguji model prediksi berbasis data mining dan machine learning. Melalui pendekatan tersebut, data yang tersedia dapat diolah menjadi informasi yang objektif dan terukur, sehingga mampu memberikan gambaran yang lebih akurat terhadap kondisi yang dianalisis. Penggunaan model prediksi juga memungkinkan pengambilan keputusan yang lebih rasional karena didasarkan pada pola yang ditemukan dalam data, bukan sekadar intuisi atau penilaian subjektif.

Dalam proses pengolahan data, terdapat tahapan penting yang dikenal dengan Knowledge Discovery in Databases (KDD). Tahapan ini merupakan rangkaian proses sistematis yang dimulai

dari pengumpulan data hingga menghasilkan pengetahuan baru yang bermanfaat. Proses KDD diawali dengan pengumpulan data dari berbagai sumber, kemudian dilanjutkan dengan tahap pembersihan data untuk menghilangkan kesalahan, duplikasi, atau data yang tidak relevan. Setelah itu, data ditransformasikan ke dalam bentuk yang sesuai agar dapat diproses lebih lanjut oleh algoritma yang digunakan.

Tahap berikutnya dalam KDD adalah penerapan algoritma data mining, seperti Naïve Bayes, yang digunakan untuk melakukan klasifikasi atau prediksi berdasarkan pola yang terdapat dalam data. Hasil dari proses ini kemudian dievaluasi untuk mengetahui tingkat akurasi dan keandalan model yang dihasilkan. Tahap akhir adalah penyajian hasil dalam bentuk informasi atau pengetahuan yang mudah dipahami dan dapat digunakan sebagai dasar dalam pengambilan keputusan. Dengan mengikuti tahapan KDD secara sistematis, proses analisis data menjadi lebih terstruktur, efektif, dan mampu menghasilkan insight yang bernilai.

7.2. Lokasi dan Subjek Kasus

Studi kasus ini dilaksanakan pada semester genap tahun ajaran 2024/2025 dengan mempertimbangkan ketersediaan data akademik yang relevan serta kebutuhan akan sistem evaluasi yang lebih efektif dan objektif. Pemilihan waktu tersebut dianggap tepat karena pada periode tersebut data siswa telah terkumpul secara lengkap, sehingga memungkinkan proses analisis dilakukan secara optimal. Selain itu, lokasi pelaksanaan studi kasus dipilih berdasarkan pertimbangan adanya kebutuhan untuk meningkatkan kualitas pengelolaan data dan sistem penilaian yang selama ini masih dilakukan secara konvensional.

Dalam studi kasus ini, populasi yang digunakan mencakup seluruh siswa yang menjadi objek pengamatan. Adapun teknik pengambilan sampel dilakukan dengan metode total sampling, sehingga seluruh anggota populasi dijadikan sampel penelitian sebanyak 60 siswa. Data yang diperoleh kemudian dibagi menjadi

dua bagian, yaitu 80% sebagai data latih (training data) yang digunakan untuk membangun model, dan 20% sebagai data uji (testing data) yang digunakan untuk mengukur kinerja dan akurasi model. Pembagian ini bertujuan agar model yang dihasilkan dapat diuji secara objektif serta mampu memberikan hasil prediksi yang lebih akurat dan dapat diandalkan.

7.3. Prosedur Penerapan Algoritma Naïve Bayes

Prosedur dalam studi kasus ini dilakukan secara sistematis untuk memastikan bahwa setiap tahapan berjalan dengan terarah dan menghasilkan keluaran yang optimal. Tahapan awal dimulai dengan studi literatur, yaitu mengumpulkan berbagai referensi yang berkaitan dengan data mining, machine learning, serta algoritma yang digunakan. Selanjutnya dilakukan identifikasi masalah untuk memahami kebutuhan dan kendala yang dihadapi, sehingga solusi yang dirancang benar-benar relevan. Tahap ini menjadi penting karena menjadi dasar dalam menentukan pendekatan dan metode yang akan digunakan dalam proses analisis.

Setelah masalah teridentifikasi, tahapan berikutnya adalah pengolahan data yang meliputi proses pembersihan, seleksi, dan transformasi data agar siap digunakan dalam pembangunan model. Selanjutnya dilakukan pembangunan model prediksi menggunakan algoritma yang telah dipilih, kemudian dilanjutkan dengan evaluasi hasil untuk mengukur tingkat akurasi dan kinerja model tersebut. Hasil evaluasi ini menjadi dasar dalam menilai apakah model yang dibangun sudah mampu memberikan prediksi yang tepat. Dengan mengikuti prosedur yang terstruktur, studi kasus ini diharapkan dapat menghasilkan sistem prediksi yang akurat dan dapat digunakan sebagai pendukung pengambilan keputusan.

7.4. Variabel Penerapan Algoritma Naïve Bayes

Dalam studi kasus ini, variabel yang digunakan terdiri dari beberapa komponen penting yang berperan dalam menggambarkan kondisi dan potensi siswa secara menyeluruh. Salah satu variabel utama adalah variabel mata pelajaran, yang mencakup tujuh mata pelajaran inti sebagai indikator kemampuan akademik siswa. Variabel ini dipilih karena nilai dari berbagai mata pelajaran dapat memberikan gambaran yang lebih komprehensif mengenai tingkat pemahaman dan pencapaian siswa dalam proses pembelajaran. Dengan mempertimbangkan lebih dari satu mata pelajaran, analisis yang dilakukan menjadi lebih objektif dan tidak hanya bergantung pada satu aspek akademik saja.

Selain variabel akademik, studi kasus ini juga mempertimbangkan variabel kehadiran sebagai indikator penting dalam proses pembelajaran. Kehadiran siswa mencerminkan tingkat partisipasi dan kedisiplinan dalam mengikuti kegiatan belajar di sekolah. Siswa dengan tingkat kehadiran yang tinggi cenderung memiliki peluang lebih besar untuk memahami materi yang diajarkan, sehingga berpotensi menunjukkan prestasi yang lebih baik. Oleh karena itu, variabel kehadiran menjadi faktor pendukung yang relevan dalam menilai dan memprediksi prestasi siswa.

Kombinasi antara variabel mata pelajaran dan kehadiran memberikan pendekatan yang lebih menyeluruh dalam membangun model prediksi. Dengan menggabungkan kedua jenis variabel tersebut, analisis tidak hanya berfokus pada hasil akademik semata, tetapi juga mempertimbangkan aspek keterlibatan siswa dalam proses pembelajaran. Pendekatan ini memungkinkan terbentuknya model prediksi yang lebih komprehensif dan akurat, karena mampu menangkap berbagai faktor yang mempengaruhi prestasi siswa secara simultan. Dengan demikian, hasil yang diperoleh dapat memberikan gambaran yang lebih tepat sebagai dasar dalam pengambilan keputusan.

Tidak untuk diperjualbelikan

7.5. Teknik Analisis Data Kasus

Dalam studi kasus ini, teknik analisis data diawali dengan tahap pengumpulan data yang menjadi fondasi utama dalam proses pengolahan. Data diperoleh melalui dokumentasi yang tersedia, berupa nilai akademik dan kehadiran siswa dengan jumlah sebanyak 50 data. Data tersebut dipilih karena memiliki keterkaitan langsung dengan indikator prestasi siswa. Proses pengumpulan data dilakukan secara sistematis untuk memastikan bahwa data yang digunakan relevan, lengkap, dan dapat mendukung proses analisis secara optimal.

Setelah data terkumpul, tahap berikutnya adalah pembersihan data. Pada tahap ini, data diperiksa untuk mengidentifikasi adanya kesalahan, duplikasi, maupun nilai yang kosong. Data yang tidak konsisten atau tidak lengkap akan diperbaiki atau dihapus agar tidak mempengaruhi hasil analisis. Proses pembersihan ini sangat penting karena kualitas data yang baik akan menghasilkan model yang lebih akurat dan dapat diandalkan.

Tahap selanjutnya adalah transformasi data, yaitu mengubah data ke dalam bentuk yang lebih sesuai untuk proses analisis. Dalam studi kasus ini, data numerik seperti nilai siswa diubah menjadi kategori, misalnya rendah, sedang, dan tinggi. Transformasi ini bertujuan untuk mempermudah proses klasifikasi, khususnya dalam penerapan algoritma yang bekerja dengan pendekatan kategorikal. Dengan adanya transformasi ini, pola dalam data menjadi lebih mudah dikenali oleh model yang digunakan.

Tahap akhir dalam teknik analisis data adalah pembagian dataset menjadi data training dan data testing. Data training digunakan untuk melatih model agar dapat mengenali pola yang terdapat dalam data, sedangkan data testing digunakan untuk menguji kinerja dan akurasi model yang telah dibangun. Pembagian ini dilakukan secara proporsional agar model dapat diuji secara objektif. Dengan melalui tahapan analisis data yang sistematis, studi kasus ini diharapkan mampu menghasilkan model prediksi

yang akurat dan dapat digunakan secara efektif dalam pengambilan keputusan.

Dilindungi oleh Undang-Undang Republik Indonesia

7.6. Pemodelan dan Evaluasi

Dalam studi kasus ini, tahap pemodelan dilakukan dengan menerapkan algoritma Naïve Bayes sebagai metode klasifikasi. Proses ini diawali dengan menghitung probabilitas prior, yaitu peluang awal suatu kelas berdasarkan data yang tersedia, serta probabilitas likelihood yang menunjukkan peluang munculnya suatu atribut terhadap kelas tertentu. Dengan menggabungkan kedua probabilitas tersebut, model dapat menentukan kemungkinan suatu data termasuk ke dalam kategori tertentu. Pendekatan probabilistik ini memungkinkan proses klasifikasi dilakukan secara sistematis dan berbasis data.

Selanjutnya, model yang telah dibangun digunakan untuk mengklasifikasikan data berdasarkan pola yang telah dipelajari dari data sebelumnya. Setiap data akan dihitung nilai probabilitasnya untuk masing-masing kelas, kemudian dikategorikan ke dalam kelas dengan probabilitas tertinggi. Proses ini menjadi inti dari penerapan Naïve Bayes karena mampu mengolah berbagai atribut secara sederhana namun tetap efektif dalam menghasilkan prediksi.

Tahap berikutnya adalah evaluasi model untuk mengetahui tingkat kinerja dari model yang telah dibangun. Evaluasi dilakukan menggunakan confusion matrix, yang memberikan gambaran mengenai hasil klasifikasi dalam bentuk perbandingan antara data aktual dan hasil prediksi. Berdasarkan evaluasi tersebut, diperoleh nilai akurasi sebesar 75%, precision sebesar 100%, recall sebesar 50%, serta F1-score sebesar 66%. Nilai-nilai ini menunjukkan bahwa model memiliki kemampuan yang baik dalam mengklasifikasikan data secara tepat terutama dalam hal precision.

Tidak untuk diperjualbelikan

Meskipun demikian, nilai recall yang masih relatif rendah menunjukkan bahwa model belum optimal dalam mendeteksi seluruh data yang seharusnya termasuk dalam kategori tertentu. Hal ini mengindikasikan bahwa masih terdapat beberapa data yang belum terklasifikasi dengan baik. Oleh karena itu, diperlukan upaya peningkatan model, seperti penambahan jumlah data, pemilihan atribut yang lebih relevan, atau penggunaan metode tambahan untuk meningkatkan performa. Dengan evaluasi yang dilakukan secara menyeluruh, model yang dihasilkan dapat terus dikembangkan agar memberikan hasil yang lebih akurat dan optimal.

7.7. Penutup

Studi kasus ini mengkaji penerapan algoritma Naïve Bayes dalam memprediksi siswa berprestasi di SMP Negeri 5 Bilah Barat. Data akademik yang digunakan meliputi nilai ujian harian, ujian tengah semester, ujian akhir semester, serta tingkat kehadiran siswa. Seluruh data tersebut telah melalui tahap preprocessing sehingga menghasilkan dataset yang terstruktur dan siap digunakan dalam proses pemodelan. Proses ini memungkinkan terbentuknya pola-pola yang relevan sebagai dasar dalam membangun model prediksi prestasi siswa secara efektif dan sistematis.

Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes dapat diterapkan dengan baik melalui tahapan pengumpulan data, preprocessing, pelatihan, hingga pengujian model. Model yang dihasilkan mampu mengklasifikasikan siswa ke dalam kategori berprestasi dan tidak berprestasi secara otomatis. Hal ini memberikan peningkatan dari segi objektivitas dan efisiensi dibandingkan metode penilaian manual. Berdasarkan pengujian menggunakan RapidMiner, model menunjukkan performa yang baik dengan nilai akurasi sebesar 90%, precision 88%, recall 87%, dan F1-score sebesar 87,5%, yang menunjukkan bahwa model cukup andal dalam melakukan prediksi.

Selain itu, studi ini juga memberikan beberapa rekomendasi untuk pengembangan penelitian selanjutnya. Disarankan untuk menambahkan variabel non-akademik seperti kegiatan ekstrakurikuler dan aspek sikap agar hasil klasifikasi menjadi lebih komprehensif. Peningkatan jumlah data juga diperlukan agar model dapat mempelajari pola yang lebih kompleks dan meningkatkan akurasi prediksi. Di samping itu, penelitian selanjutnya dapat melakukan perbandingan dengan metode lain serta mengembangkan sistem berbasis web atau mobile, sehingga hasil penelitian dapat diimplementasikan secara langsung dan lebih optimal dalam mendukung pengambilan keputusan di lingkungan sekolah.



Penerapan metode machine learning, khususnya algoritma Naïve Bayes, memberikan kontribusi yang signifikan dalam meningkatkan kualitas proses identifikasi siswa berprestasi. Pendekatan berbasis data ini memungkinkan proses klasifikasi dilakukan secara objektif, sistematis, dan terukur, sehingga mampu mengurangi bias yang sering muncul dalam metode penilaian konvensional.

Pengolahan data akademik seperti nilai mata pelajaran, tingkat kehadiran, serta atribut pendukung lainnya memungkinkan terbentuknya pola-pola yang mencerminkan kinerja siswa. Pola tersebut kemudian digunakan sebagai dasar dalam membangun model prediksi yang akurat. Dengan memanfaatkan data yang tersedia secara optimal, pihak sekolah dapat memperoleh gambaran yang lebih jelas mengenai potensi dan perkembangan siswa, sehingga proses evaluasi menjadi lebih terarah dan efektif.

Selain itu, penerapan algoritma Naïve Bayes juga memberikan kemudahan dalam implementasi serta efisiensi dalam proses komputasi. Hal ini menjadikannya sebagai solusi yang tepat, terutama bagi institusi pendidikan yang ingin mengadopsi

teknologi tanpa memerlukan sumber daya yang besar. Dengan demikian, pemanfaatan machine learning dalam analisis data akademik tidak hanya meningkatkan kualitas penilaian, tetapi juga mendukung pengambilan keputusan yang lebih tepat dalam upaya meningkatkan prestasi siswa secara keseluruhan.

Penerapan algoritma Naïve Bayes terbukti mampu mengklasifikasikan siswa ke dalam kategori berprestasi dan tidak berprestasi dengan tingkat kinerja yang baik. Keunggulan algoritma ini terletak pada kesederhanaan perhitungan, efisiensi komputasi, serta kemampuannya dalam bekerja pada dataset berukuran kecil hingga menengah. Hal ini menjadikannya sebagai solusi yang tepat, khususnya bagi institusi pendidikan yang memiliki keterbatasan dalam infrastruktur teknologi namun tetap ingin memanfaatkan analisis data dalam pengambilan keputusan.

Selain itu, penggunaan metrik evaluasi seperti akurasi, precision, recall, dan F1-score memberikan gambaran yang komprehensif mengenai performa model yang dibangun. Evaluasi yang tepat memastikan bahwa model tidak hanya memiliki tingkat ketepatan yang tinggi, tetapi juga mampu mendeteksi siswa yang berpotensi mengalami penurunan prestasi. Dengan demikian, hasil prediksi dapat digunakan sebagai dasar dalam merancang strategi pembelajaran dan intervensi yang lebih efektif.

Secara keseluruhan, penerapan algoritma Naïve Bayes dalam konteks pendidikan memberikan dampak yang signifikan terhadap peningkatan kualitas proses evaluasi akademik. Dengan memanfaatkan pendekatan berbasis data, penilaian siswa tidak lagi bergantung pada metode manual yang cenderung subjektif, melainkan didasarkan pada analisis yang sistematis dan terukur. Hal ini memungkinkan sekolah untuk memperoleh hasil evaluasi yang lebih akurat serta mencerminkan kondisi nyata dari kinerja siswa.

Selain itu, penggunaan algoritma ini juga mendorong transformasi menuju sistem pendidikan berbasis data. Sekolah dapat mengelola dan menganalisis data akademik secara lebih efektif, sehingga

mampu mengidentifikasi pola-pola penting yang berkaitan dengan prestasi siswa. Transformasi ini menjadi langkah penting dalam menghadapi perkembangan teknologi, di mana pengambilan keputusan berbasis data menjadi kebutuhan utama dalam meningkatkan kualitas pendidikan.

Dengan demikian, pendekatan ini diharapkan dapat membantu pihak sekolah dalam mengoptimalkan potensi siswa secara lebih maksimal. Selain meningkatkan objektivitas dalam penilaian, hasil analisis juga dapat digunakan sebagai dasar dalam merancang strategi pembelajaran dan intervensi yang lebih tepat. Dalam jangka panjang, penerapan teknologi seperti Naïve Bayes akan mendukung terciptanya sistem pendidikan yang lebih adaptif, efisien, dan berkelanjutan.

©p.pustakariyadz

Dilindungi oleh Undang-Undang Republik Indonesia
Nomor 28 Tahun 2014 tentang Hak Cipta

©p.pustakariyadz



DAFTAR PUSTAKA

- D., P. and-. (2025a). Scaling up data mining algorithms for big data. *International Journal for Multidisciplinary Research*, 7(1).
<https://doi.org/https://doi.org/10.36948/ijfmr.2025.v07i01.34838>
- P., & -, D. (2025). Scaling up data mining algorithms for big data. *International Journal for Multidisciplinary Research*, 7(1).
<https://doi.org/https://doi.org/10.36948/ijfmr.2025.v07i01.34838>
- (2020). ANN and SVM Algorithm in Divorce Predictor. *International Journal of Engineering and Advanced Technology*.
<https://doi.org/10.35940/ijeat.c5902.029320>. (n.d.).
- (2023). (2023). Texture features extraction technology using grey level co-occurrence matrix for the knn classification of citrus disease. *Arpn Journal of Engineering and Applied Sciences*, 919-925. <https://doi.org/10.59018/0423122>
- Adane Deku, J., & Asare, E., M. (2023). Performance Analysis of Machine Learning Algorithms in Prediction of Student Academic Performance. *Journal of Advances in Mathematics and Computer Science*. <https://doi.org/10.9734/jamcs/2023/v38i51762>
- Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., & Shafique, M. (2022). Predicting at-risk students using machine learning techniques. *IEEE Access*, 10, 6735–6746.
<https://doi.org/10.1109/ACCESS.2022.3140281>

- Afridah Ula, M., & Rosnita, L. R. (2024). (2024). *Performance Analysis Algorithm Classification and Regression Trees and Naive Bayes Based Particle Swarm Optimization for Credit Card Transaction Fraud Detection*. <https://doi.org/10.52088/ijesty.v4i3.523>
- Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer. <https://doi.org/10.1007/978-3-319-14142-8>
- Ahaggach, H., Abrouk, L., & Lebon, E. (2024). Systematic mapping study of sales forecasting: Methods, trends, and future directions. *Forecasting*. <https://doi.org/https://doi.org/10.3390/forecast6030028>
- Ahalt Avillach, P., Boyles, R., Bradford, K., Cox, S., Davis-Dusenbery, B., ... & Asare, J. S. (2023). (2023). *Building a Collaborative Cloud Platform to Accelerate Heart, Lung, Blood, and Sleep Research*. <https://doi.org/10.1093/jamia/ocad048>
- Ahmed, E. (2024). Student Performance Prediction Using Machine Learning Algorithms. *Appl. Comput. Intell. Soft Comput.*, 4067721:1-4067721:15. <https://doi.org/10.1155/2024/4067721>
- Ahsan, M., Mahmud, M., Saha, P., Gupta, K., & Siddique, Z. (2021). *Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance*. *Technologies*. <https://doi.org/10.3390/technologies9030052>. (n.d.).
- Aji, G., & Devi, P. (2023). Data Mining Implementation for Product Transaction Patterns Using Apriori Method. *Sinkron*. <https://doi.org/https://doi.org/10.33395/sinkron.v8i1.12071>
- Akpınar, C., Koşar, Ö., & Durdu, A. (2025). *Enhancing the Performance of Machine Learning Classification Models*. *2025 24th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 1-6. <https://doi.org/10.1109/infotech64129.2025.10959301>. (n.d.).
- Al-Barrak, M. A., & Al-Razgan, M. S. (2016). Predicting students final GPA using decision trees: A case study. *International Journal of Information and Education Technology*, 6(7), 528–533. <https://doi.org/10.7763/IJiet.2016.V6.745>

- Alam, S., & Yao, N. (2018). *The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. Computational and Mathematical Organization Theory*, 25, 319 - 335. <https://doi.org/10.1007/s10588-018-9266-8>. (n.d.).
- Ali, R., Zada, M., Khattak, A., & Hussain, J. (2023). Algorithm selection using edge ML and case based reasoning. *Journal of Cloud Computing*, 12, 1-19. <https://doi.org/10.1186/s13677-023-00542-3>
- Alsariera Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A., & Ali, N., Y. (2022). Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance. *Computational Intelligence and Neuroscience*, 2022. <https://doi.org/10.1155/2022/4151487>
- Amato, P. (2010). Research on divorce: Continuing trends and new developments. *Journal of Marriage and Family*, 72, 650-666. <https://doi.org/10.1111/j.1741-3737.2010.00723.x>
- Amornsamankul, S., Pimpunchat, B., Triampo, W., Charoenpong, J., & Nuttavut, N. (2019). A comparison of machine learning algorithms and their applications. *International Journal of Simulation: Systems, Science & Technology*. <https://doi.org/10.5013/ijssst.a.20.04.08>
- Aridas Karlos, S., Kanas, V., Fazakis, N., & Kotsiantis, S, C. (n.d.). (2020). *Uncertainty Based Under-Sampling for Learning Naive Bayes Classifiers Under Imbalanced Data Sets*, IEEE Access-8. <https://doi.org/https://doi.org/10.1109/access.2019.2961784>
- Arora Pathak, S., & Linh, N, K. (2023). (2023). *Comparative Analysis Ofk-Nn, Naïve Bayes, and Logistic Regression for Credit Card Fraud Detection*. <https://doi.org/10.16925/2357-6014.2023.03.05>
- Arpino, B., Moglie, M., & Mencarini, L. (2021). *What Tears Couples Apart: A Machine Learning Analysis of Union Dissolution in Germany.. Demography*. <https://doi.org/10.1215/00703370-9648346>. (n.d.).

- Asai Tashiro, T., Kondo, Y., Hayashi, M., Arihara, H., Omote, S., ... & Yamamoto, Y. Y. (2023). (2023). *Machine Learning-Based Prediction of Digoxin Toxicity in Heart Failure: A Multicenter Retrospective Study*. <https://doi.org/10.1248/bpb.b22-00823>
- Ashraf Khalilitousi, M., & Laksman, Z, M. (2021). (2021). *Applying Machine Learning to Stem Cell Culture and Differentiation*. <https://doi.org/10.1002/cpz1.261>
- Atmaja Nugraha, I., & Ambaradewi, N, I. (2023). (2023). *Data Mining Memprediksi Kelulusan Mahasiswa Menggunakan Metode K-Nearest Neighbors (Knn) Studi Kasus Universitas Pgri Mahadewa Indonesia*. <https://doi.org/10.59819/jmti.v13i2.3082>
- Azad, M. (n.d.). A. *Comparative Study of Naïve Bayes and KNN Algorithms for Classification Problems, International Journal of Computer Applications*, 185(15), 1–7.).
- Aziz, A., & Rofik, R. (2025). Comparison of naïve bayes and support vector machine methods for jkt48 music video comment classification. *Journal of Student Research Exploration*. <https://doi.org/10.52465/josre.v3i1.389>
- Baker, R. S. J. d., & Inventado, P. S. (2014). *Educational data mining and learning analytics*. In *Learning analytics* (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4
- Basuki Nurfauzi, A, A. (2023). R., & Ma'arif, M. F. [https://eprints.uad.ac.id/43358/1/17-Sentiment Analysis of Customers' Review on Delivery Service Provider on Twitter Using Naive Bayes Classification.pdf](https://eprints.uad.ac.id/43358/1/17-Sentiment%20Analysis%20of%20Customers'%20Review%20on%20Delivery%20Service%20Provider%20on%20Twitter%20Using%20Naive%20Bayes%20Classification.pdf)
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281–305.
- Bhardwaj, B. K., & Pal, S. (2019). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications*, 10(5), 186–193.
- Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., & Sillero-Denamiel, M. (2021). Variable selection for Naïve Bayes classification.

- Computers & Operations Research*, 133, 105456. <https://doi.org/https://doi.org/10.1016/j.cor.2021.105456>
- Dilindungi oleh Undang-Undang Republik Indonesia
- Bonab Tanha, J., & Masdari, M, M. (2024). (2024). *A Semi-Supervised Learning Approach to Quality-Based Web Service Classification*. <https://doi.org/10.1109/access.2024.3385341>
- Boubaker Le, T., Ngo, T., & Manita, R, S. (n.d.). (2023). *Predicting the Performance of MSMEs: A Hybrid DEA-Machine Learning Approach*, *Annals of Operations Research-1-23*. <https://doi.org/https://doi.org/10.1007/s10479-023-05230-8>
- Bradley, A. (1997). *The use of the area under the ROC curve in the evaluation of machine learning algorithms*. *Pattern Recognit.*, 30, 1145-1159. [https://doi.org/10.1016/s0031-3203\(96\)00142-2](https://doi.org/10.1016/s0031-3203(96)00142-2). (n.d.).
- Bravo-Rocca Liu, P., Guitart, J., Dholakia, A., Ellison, D., Falkanger, J., ... & Hodak, M, G. (2022). (2022). *Scanflow: A Multi-Graph Framework for Machine Learning Workflow Management, Supervision, and Debugging*. <https://doi.org/10.1016/j.eswa.2022.117232>
- Celeste, J., Tasnim, M., Cuervo, A., Cal, E., & Stroulia, E. (2025). A software pipeline for systematizing machine learning of speech data. *Frontiers in Psychiatry*, 16. <https://doi.org/https://doi.org/10.3389/fpsy.2025.1451368>
- Cervantes, J., García, F., Rodríguez-Mazahua, L., & López-Chau, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408, 189–215. <https://doi.org/10.1016/j.neucom.2019.10.113>
- Chhimwal, R., & Sinha, G. R. (2025). Handling class imbalance in educational data mining: A machine learning perspective. *International Journal of Learning Analytics and Artificial Intelligence*, 4(1), 1–14.

- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *Proceedings of the 5th Future Business Technology Conference*, 5–12.
- Costa-Climent Haftor, D., & Staniewski, M, R. (n.d.). (2023). *Using Machine Learning to Create and Capture Value in the Business Models of Small and Medium-Sized Enterprises*, International Journal of Information Management-73. <https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2023.102637>
- Cui, S. (2025). Classification of daily quantities of emergency cases based on Bayesian discriminant analysis. *Proceedings of the 2nd International Conference on Modeling, NLP and Machine Learning*. <https://doi.org/https://doi.org/10.1145/3757110.3757128>
- Cunningham & Delany, S, P. (n.d.). *J. K-Nearest Neighbour Classifiers, 2020*, Multiple Classifier Systems-34.
- Dalir, Z., Nedaei, B., & Arian, M. (2025). Machine learning algorithms for predicting the impact of care burden on the psychological well-being of caregivers for chronic kidney disease patients. *Iranian Journal of Nursing and Midwifery Research*, 30(5), 682–691. https://doi.org/https://doi.org/10.4103/ijnmr.ijnmr_393_23
- Dauner Leal, E., Adam, T., Zhang, R., & Farley, J., D. (2023). Evaluation of four machine learning models for signal detection. *Therapeutic Advances in Drug Safety*, 14. <https://doi.org/https://doi.org/10.1177/20420986231219472>
- Debora Pratama, I., & Sasmita, G., U. (2024). Sentiment analysis of domestic violence issues on twitter using multinomial naïve bayes and support vector machine. *Jitter Jurnal Ilmiah Teknologi Dan Komputer*, 4(3), 1992. <https://doi.org/https://doi.org/10.24843/jtrti.2023.v04.i03.p06>

- Verma Yadav, R., & Kholiya, K., S. (2022). Prediction of Academic Performance of Engineering Students by Using Data Mining Techniques. *International Journal of Information and Education Technology*. <https://doi.org/10.18178/ijiet.2022.12.11.1734>
- Zheng & Li, C., X. (2024). Predicting Students' Academic Performance Through Machine Learning Classifiers: A Study Employing the Naive Bayes Classifier (NBC). *International Journal of Advanced Computer Science and Applications*. <https://doi.org/10.14569/ijacsa.2024.0150199>

©p.pustakariyadz